Electronics and
Computer Science

UNIVERSITY OF
Southampton

Introduction to Information Extraction

Dr. Elena Demidova

University of Southampton

ESWC Summer School 2016

September 6, 2016

Vision

- Information related to real-world entities, their relations, events and topics is currently scattered across different sources on the Web, in Web archives, digital libraries, news sources, open datasets, Linked Data and Social Media.
- This information should evolve in a virtual informational space, providing different e.g. language-specific views, provenance and temporal context.

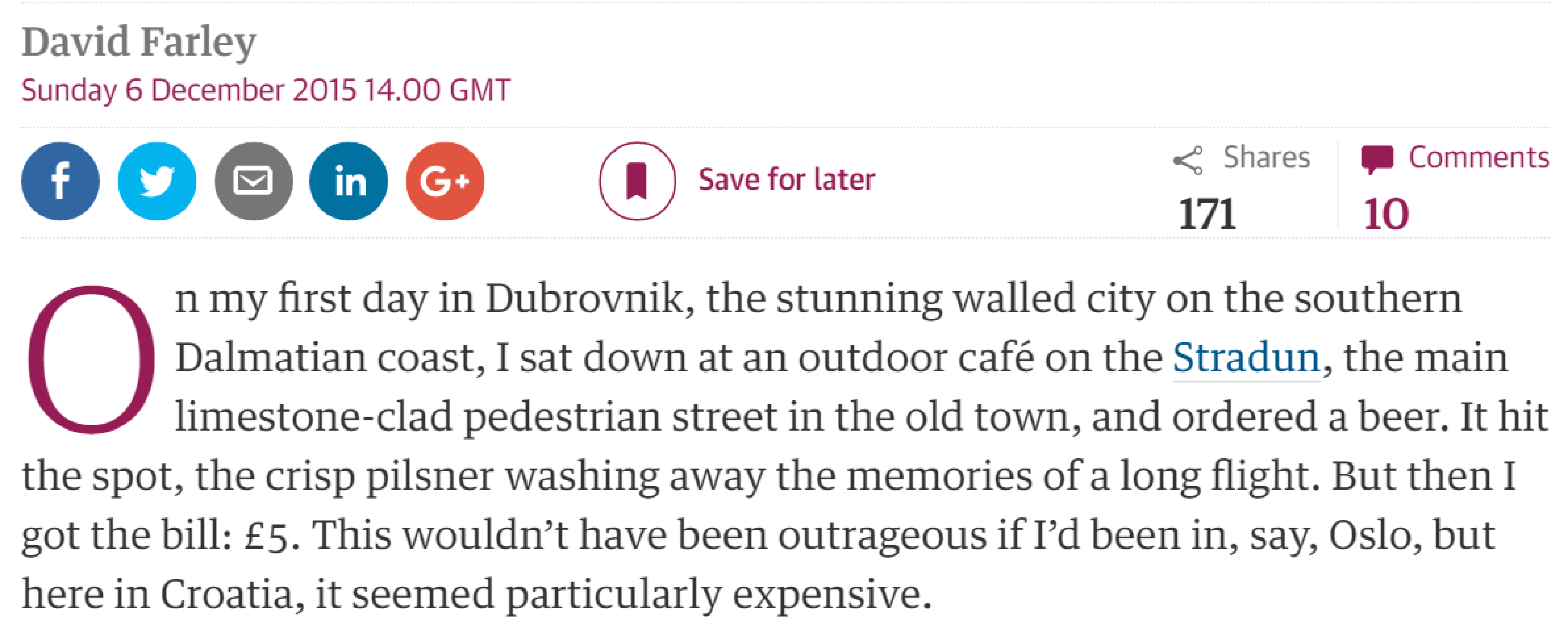
Motivation

- The amount of unstructured information on the Web is ever growing (the Web, news and social media streams, Web archives, etc.).
- Unstructured data, i.e. text, is written for humans, not machines.
- Information Extraction enables to automatically identify information nuggets such as named entities, time expressions, relations and events in text and interlink these information nuggets with structured background knowledge.
- Extracted information can then be used e.g. to categorize and cluster text, enable faceted exploration, extract semantics, populate knowledge bases, correlate extracted data with other sources, e.g. across languages etc.

Contents

- Information Extraction
- Named Entity Extraction
- Named Entity Linking
- Temporal Extraction
- Relation Extraction

(Semi-)structured Information on the Web



David Farley

Sunday 6 December 2015 14.00 GMT

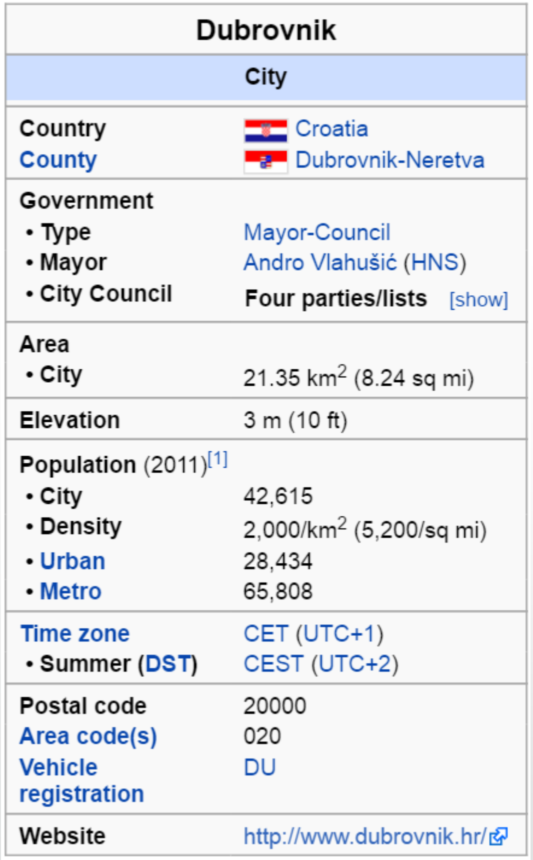
Save for later

Shares 171

Comments 10

On my first day in Dubrovnik, the stunning walled city on the southern Dalmatian coast, I sat down at an outdoor café on the Stradun, the main limestone-clad pedestrian street in the old town, and ordered a beer. It hit the spot, the crisp pilsner washing away the memories of a long flight. But then I got the bill: £5. This wouldn't have been outrageous if I'd been in, say, Oslo, but here in Croatia, it seemed particularly expensive.

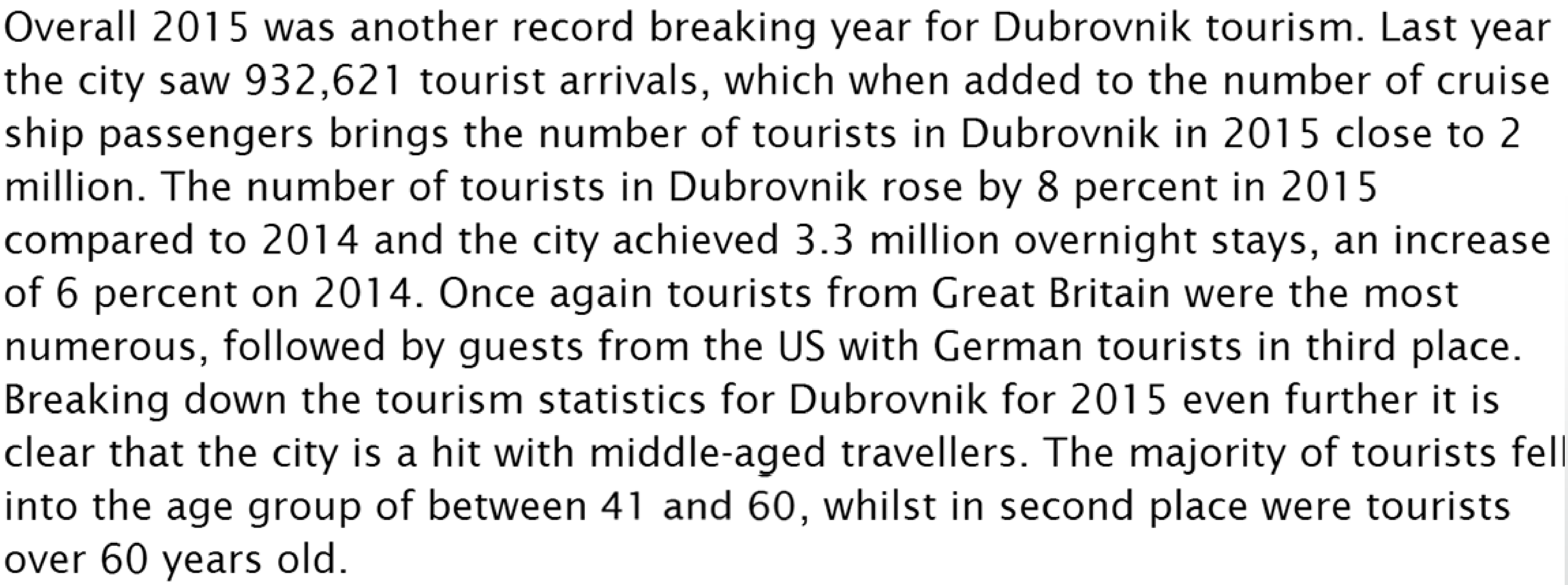
https://www.theguardian.com/travel/2015/dec/06/bar-tour-dubrovnik-croatia-holiday



Dubrovnik	
City	
Country	Croatia
County	Dubrovnik-Neretva
Government	
• Type	Mayor-Council
• Mayor	Andro Vlahušić (HNS)
• City Council	Four parties/lists [show]
Area	
• City	21.35 km^2 (8.24 sq mi)
Elevation	3 m (10 ft)
Population (2011)[1]	
• City	42,615
• Density	2,000/km^2 (5,200/sq mi)
• Urban	28,434
• Metro	65,808
Time zone	CET (UTC+1)
• Summer (DST)	CEST (UTC+2)
Postal code	20000
Area code(s)	020
Vehicle registration	DU
Website	http://www.dubrovnik.hr/

https://en.wikipedia.org/wiki/Dubrovnik

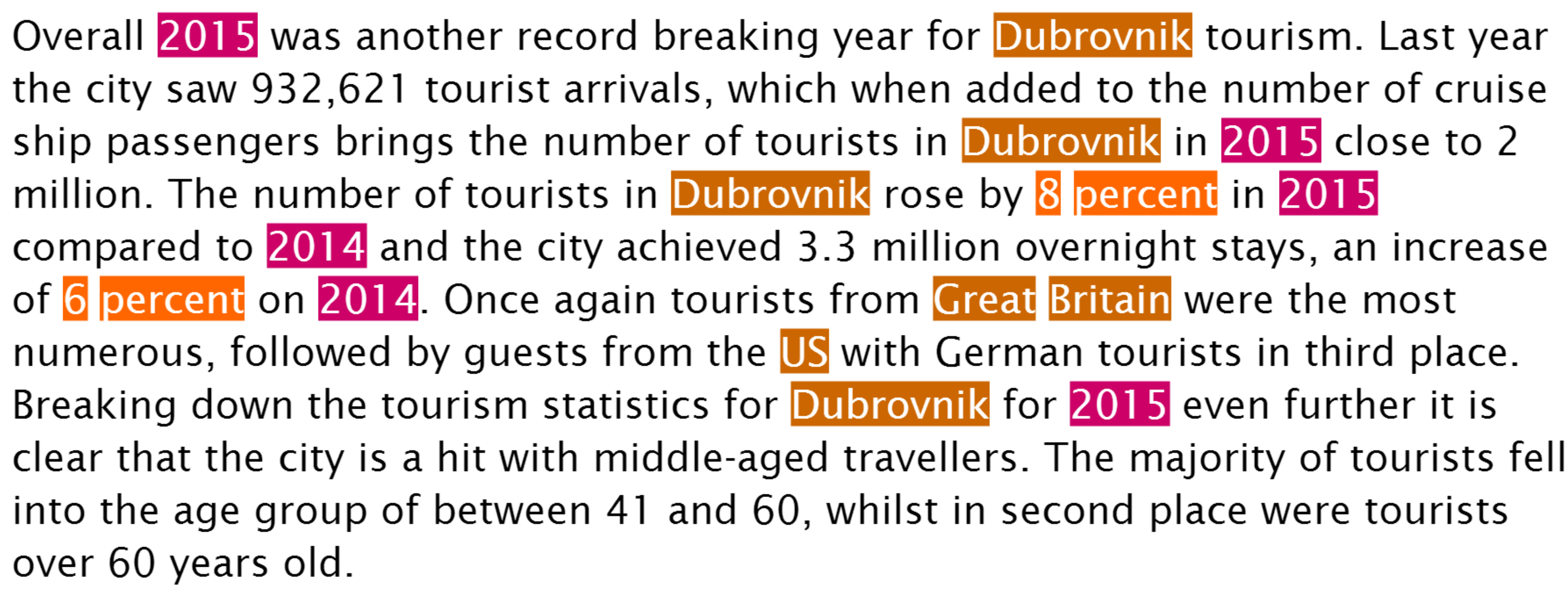
Data “Hidden” in the Text



Overall 2015 was another record breaking year for Dubrovnik tourism. Last year the city saw 932,621 tourist arrivals, which when added to the number of cruise ship passengers brings the number of tourists in Dubrovnik in 2015 close to 2 million. The number of tourists in Dubrovnik rose by 8 percent in 2015 compared to 2014 and the city achieved 3.3 million overnight stays, an increase of 6 percent on 2014. Once again tourists from Great Britain were the most numerous, followed by guests from the US with German tourists in third place. Breaking down the tourism statistics for Dubrovnik for 2015 even further it is clear that the city is a hit with middle-aged travellers. The majority of tourists fell into the age group of between 41 and 60, whilst in second place were tourists over 60 years old.

http://dubrovacki.hr/clanak/81000/2015-tourism-figures-for-dubrovnik

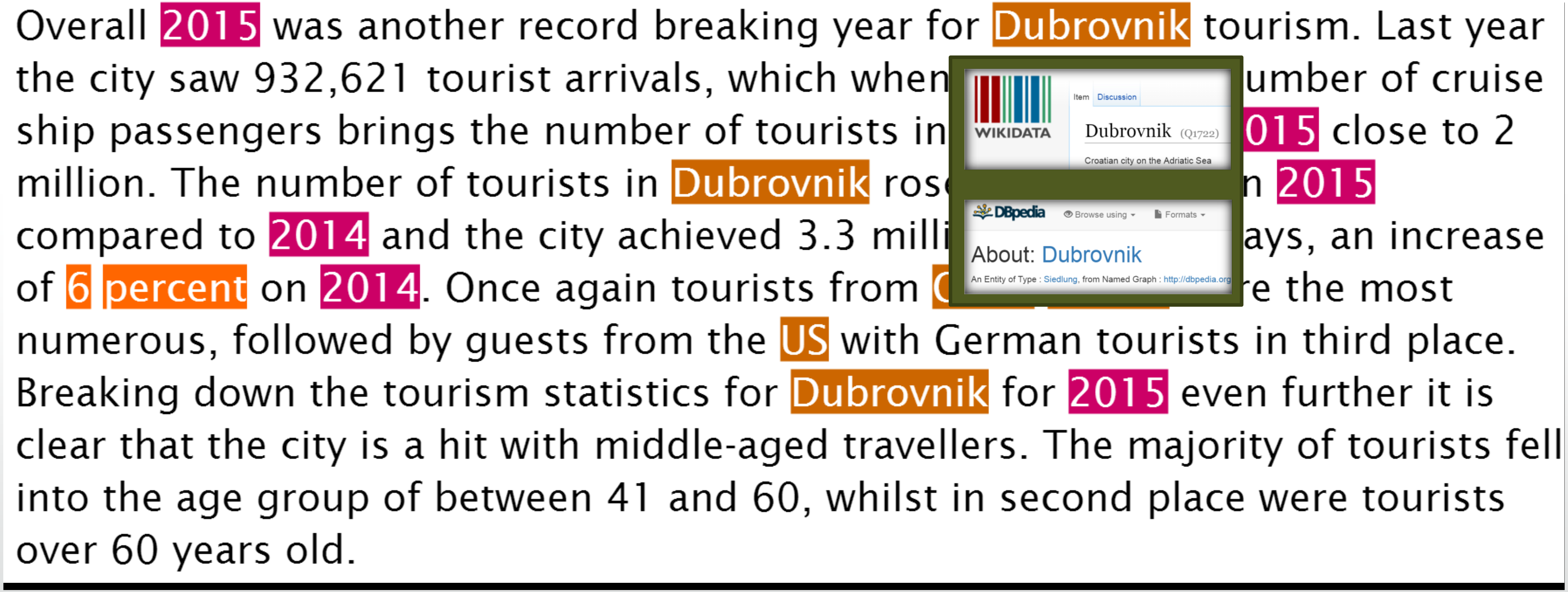
Named Entities & Temporal Expressions



Overall 2015 was another record breaking year for Dubrovnik tourism. Last year the city saw 932,621 tourist arrivals, which when added to the number of cruise ship passengers brings the number of tourists in Dubrovnik in 2015 close to 2 million. The number of tourists in Dubrovnik rose by 8 percent in 2015 compared to 2014 and the city achieved 3.3 million overnight stays, an increase of 6 percent on 2014. Once again tourists from Great Britain were the most numerous, followed by guests from the US with German tourists in third place. Breaking down the tourism statistics for Dubrovnik for 2015 even further it is clear that the city is a hit with middle-aged travellers. The majority of tourists fell into the age group of between 41 and 60, whilst in second place were tourists over 60 years old.

http://dubrovacki.hr/clanak/81000/2015-tourism-figures-for-dubrovnik

Entity Linking



Overall 2015 was another record breaking year for Dubrovnik tourism. Last year
the city saw 932,621 tourist arrivals, which when umber of cruise
ship passengers brings the number of tourists in 015 close to 2
million. The number of tourists in Dubrovnik ros n 2015
compared to 2014 and the city achieved 3.3 milli ays, an increase
of 6 percent on 2014. Once again tourists from C e the most
numerous, followed by guests from the US with German tourists in third place.
Breaking down the tourism statistics for Dubrovnik for 2015 even further it is
clear that the city is a hit with middle-aged travellers. The majority of tourists fell
into the age group of between 41 and 60, whilst in second place were tourists
over 60 years old.

WIKIDATA

Item Discussion

Dubrovnik (Q1722)

Croatian city on the Adriatic Sea

DBpedia

Browse using ▾ Formats ▾

About: Dubrovnik

An Entity of Type : Siedlung, from Named Graph : http://dbpedia.org

http://dubrovacki.hr/clanak/81000/2015-tourism-figures-for-dubrovnik

Information Extraction (IE)

- IE is the task of identification of structured information in text. IE includes:
  - **Named Entity extraction and disambiguation**
    - Dubrovnik is a city. Dubrovnik ->http://dbpedia.org/resource/Dubrovnik
  - **Extraction of temporal expressions**
    - 10th July, 25th August 2016.
  - **Extraction of relations between Named Entities**
    - Dubrovnik is located in the region of Dalmatia.
  - **Event extraction**
    - One of Croatia's most famous events, the Dubrovnik Summer Festival, took place from 10th July to 25th August 2016.

Named Entity Extraction: Terminology

Named Entities: Proper nouns or phrases, which refer to real-world objects (entities).

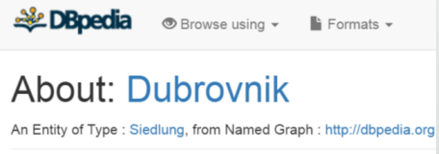
Named Entity Extraction (Recognition, Identification): Detecting boundaries of named entities (NEs) in unstructured text.

Named Entity Classification: Automatically assigning pre-defined classes to NEs, such as PERSON, LOCATION, ORGANISATION, etc.

Named Entity Linking / Disambiguation: Linking NEs to entries in a knowledge base (e.g. DBpedia, Wikidata, etc.):

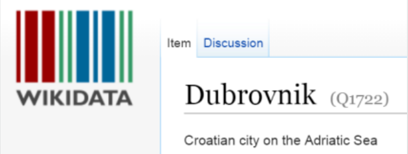
Dubrovnik -> http://dbpedia.org/resource/Dubrovnik

-> https://www.wikidata.org/wiki/Q1722



About: Dubrovnik

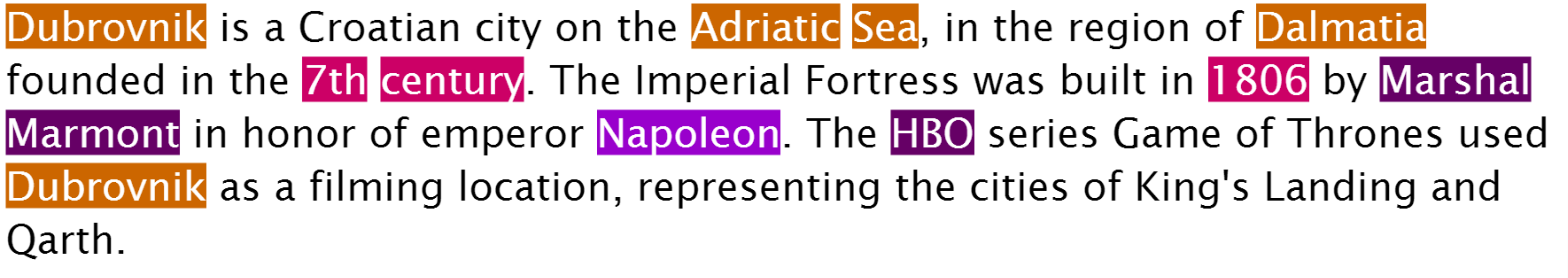
An Entity of Type : Siedlung, from Named Graph : http://dbpedia.org



Dubrovnik (Q1722)

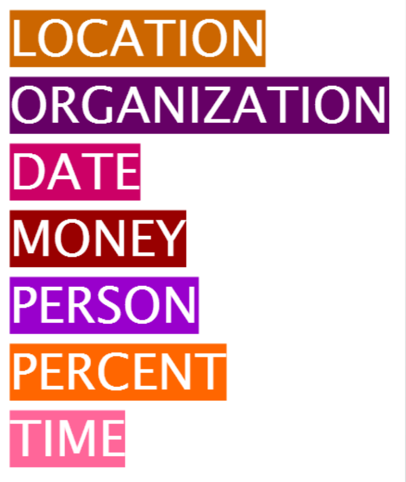
Croatian city on the Adriatic Sea

Named Entity Extraction: Examples



Dubrovnik is a Croatian city on the Adriatic Sea, in the region of Dalmatia founded in the 7th century. The Imperial Fortress was built in 1806 by Marshal Marmont in honor of emperor Napoleon. The HBO series Game of Thrones used Dubrovnik as a filming location, representing the cities of King's Landing and Qarth.

NE Classification

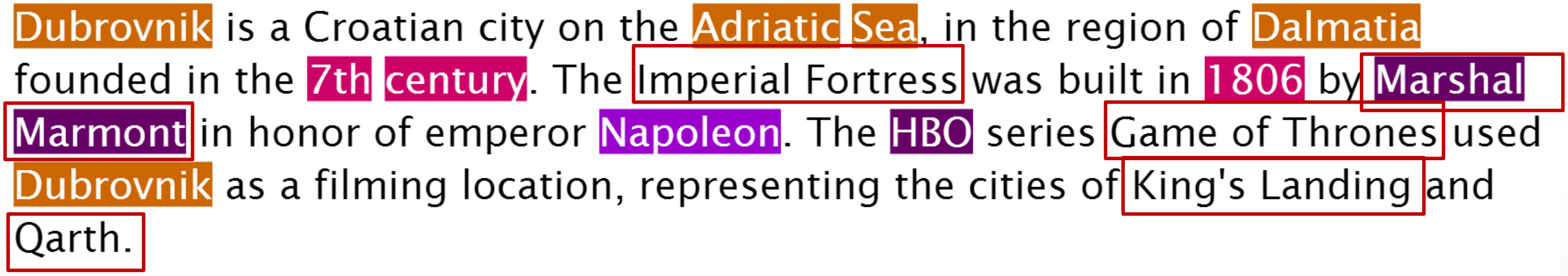


LOCATION
ORGANIZATION
DATE
MONEY
PERSON
PERCENT
TIME

Extraction by Stanford Named Entity Tagger
http://nlp.stanford.edu:8080/ner/process

Any issues?

Named Entity Extraction: Examples

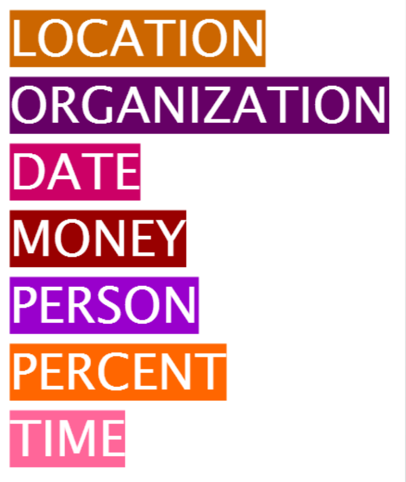


Dubrovnik is a Croatian city on the Adriatic Sea, in the region of Dalmatia founded in the 7th century. The Imperial Fortress was built in 1806 by Marshal Marmont in honor of emperor Napoleon. The HBO series Game of Thrones used Dubrovnik as a filming location, representing the cities of King's Landing and Qarth.

Problems:

- Unknown entities
- Unknown entity types
- Ambiguities / wrong types

NE Classification

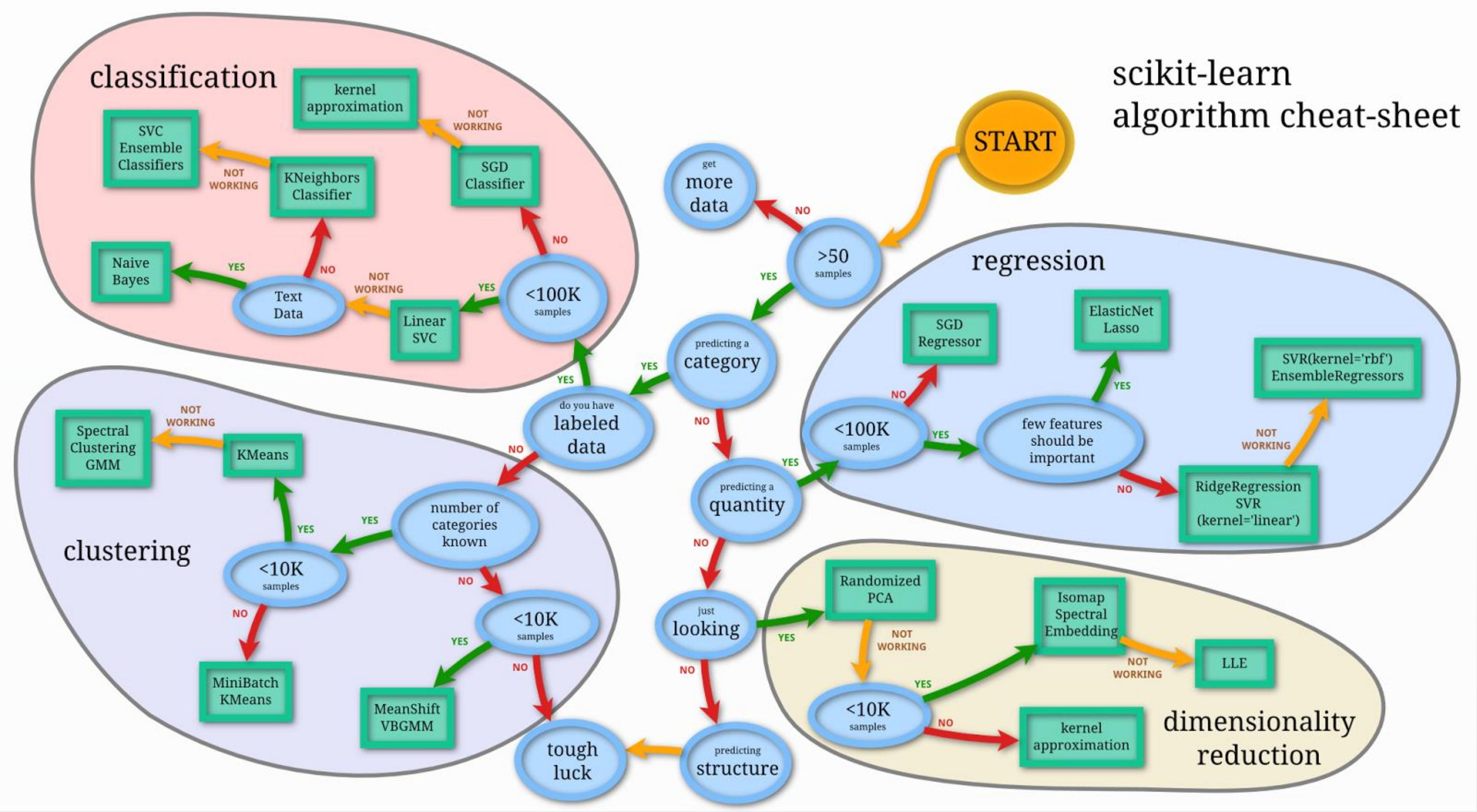


LOCATION
ORGANIZATION
DATE
MONEY
PERSON
PERCENT
TIME

Named Entity Extraction: Methods

- **Rule-based approaches**: Using hand-coded extraction rules
- **Machine learning based approaches**
  - Supervised learning (domain specific): Manually annotate the text, train a model
  - Unsupervised learning (Web-scale NER): Extract language patterns, cluster similar ones
  - Semi-supervised learning: Start with a small number of language patterns, iteratively learn more (bootstrapping)
- **Methods based on existing resources**
  - Gazetteer-based method: Use existing list of named entities
  - Using Web resources and KBs: Wikipedia, DBpedia, Web n-gram corpora, etc.
- **Combinations of the methods above**

NERC: Choice of Machine Learning Algorithms



NE Extraction Pipeline

- **Pre-processing of text**
  - Text extraction (mark up removal), sentence splitting, tokenization (identification of individual terms)
- **Linguistic pre-processing of tokens**
  - Lemmatisation (lexicon) or stemming (algorithms):
    - reduce inflectional forms of a word to a common base form
  - Part of speech (POS) tagging
- **Chunking (shallow parsing), parsing (parse tree)**
  - Noun phrases, grammatical relations
- **Semantic and discourse analysis, anaphora resolution (co-references)**
  - What actions are being described? What roles entities play in this actions? How does they relate to other entities and actions in other sentences?

NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia*.

NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia.*
- Tokenization: Dubrovnik is located in the region of Dalmatia.

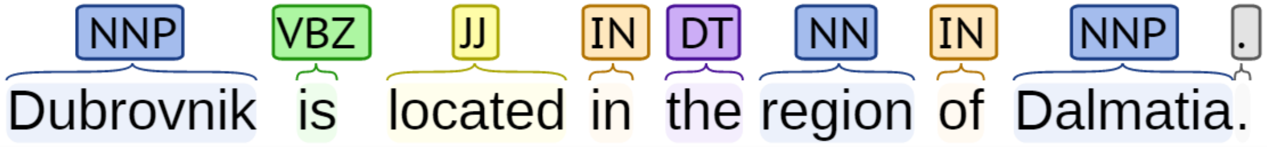
杜布羅夫尼克位於達爾馬提亞地區。

NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia*.
- Tokenization: Dubrovnik is located in the region of Dalmatia.
- Lemmatisation or stemming (reduce inflectional forms of a word to a common base form):
  - E.g. “located” -> “locate”

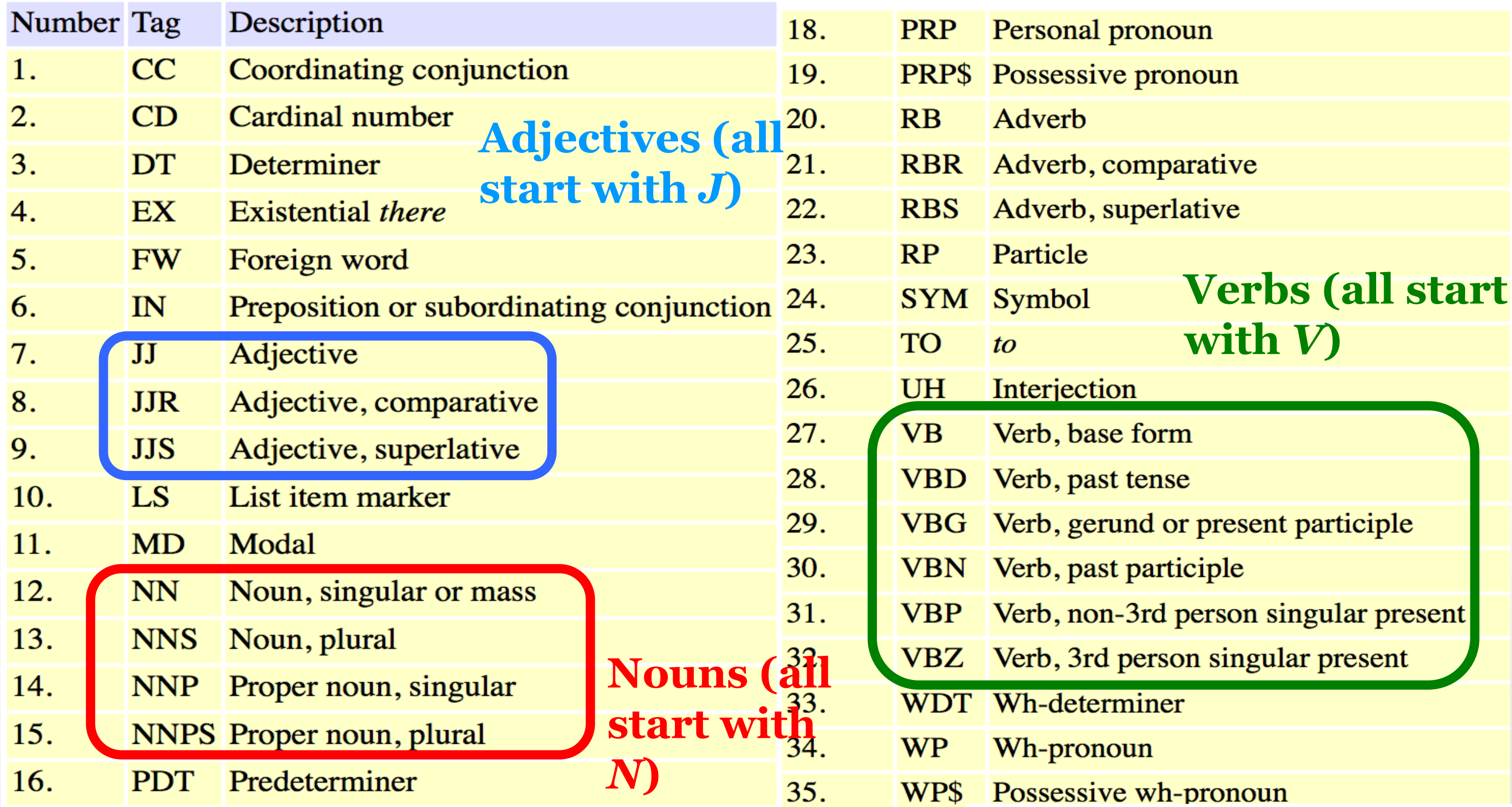
NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia.*
- Tokenization: Dubrovnik is located in the region of Dalmatia.
- Lemmatisation or stemming:
  - E.g. “located” -> “locate”



- POS tagging: Nouns, adjectives and verbs

Morphology: Penn Treebank POS tags

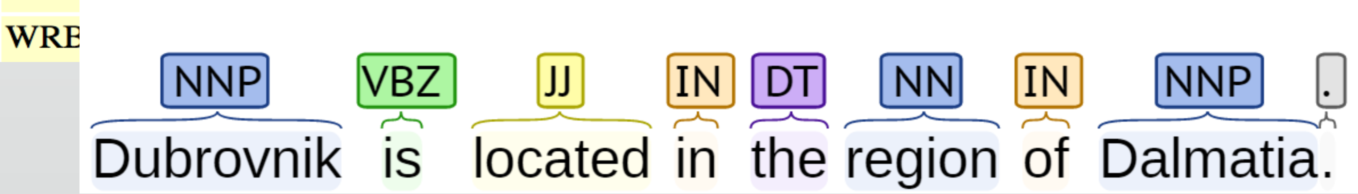


Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	

Adjectives (all start with *J*)

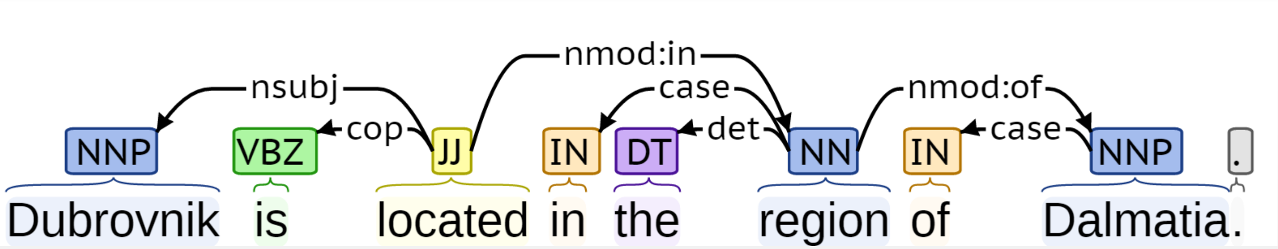
Nouns (all start with *N*)

Verbs (all start with *V*)



NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia.*
- Tokenization: Dubrovnik is located in the region of Dalmatia.
- Lemmatisation or stemming:
  - E.g. “located” -> “locate”
- POS tagging: Nouns, adjectives and verbs
- Chunking, parsing
  - “Dubrovnik is located”, “region of Dalmatia”

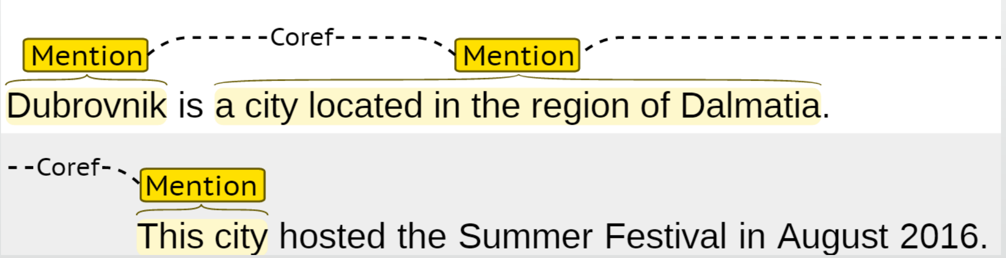


Noun phrases, grammatical relations

nsubj: nominal subject. A nominal subject is a noun phrase which is the syntactic subject of a clause.

NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia.*
- Tokenization: Dubrovnik is located in the region of Dalmatia.
- Lemmatisation or stemming:
  - E.g. “located” -> “locate”
- POS tagging: Nouns, adjectives and verbs
- Chunking, parsing
  - “Dubrovnik is located”, “region of Dalmatia”
- Co-reference resolution
  - “Dubrovnik” and “This city”.
  - Examples: http://nlp.stanford.edu:8080/corenlp/process

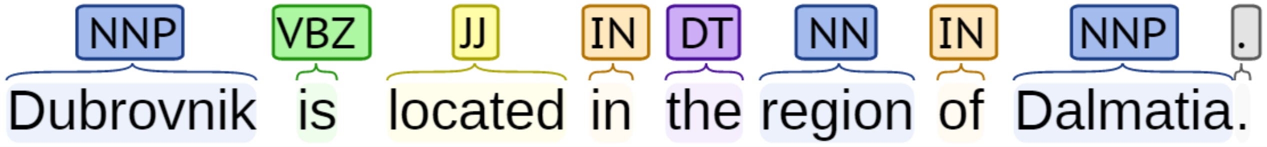


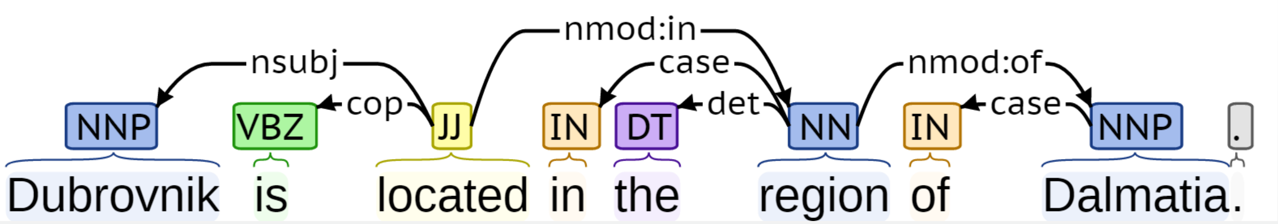
Dubrovnik is a city located in the region of Dalmatia.

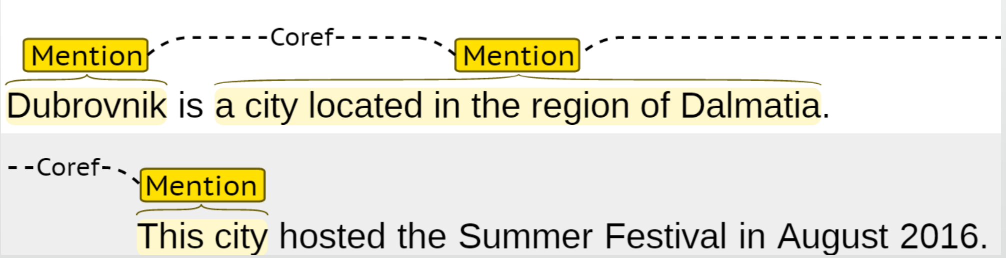
This city hosted the Summer Festival in August 2016.

NE Extraction Pipeline

- Sentence splitting: *Dubrovnik is located in the region of Dalmatia.*
- Tokenization: Dubrovnik is located in the region of Dalmatia.
- Lemmatisation or stemming:
  - E.g. “located” -> “locate”
- POS tagging: Nouns, adjectives and verbs
- Chunking, parsing
  - “Dubrovnik is located”, “region of Dalmatia”
- Co-reference resolution
  - “Dubrovnik” and “This city”.
  - Examples: http://nlp.stanford.edu:8080/corenlp/process





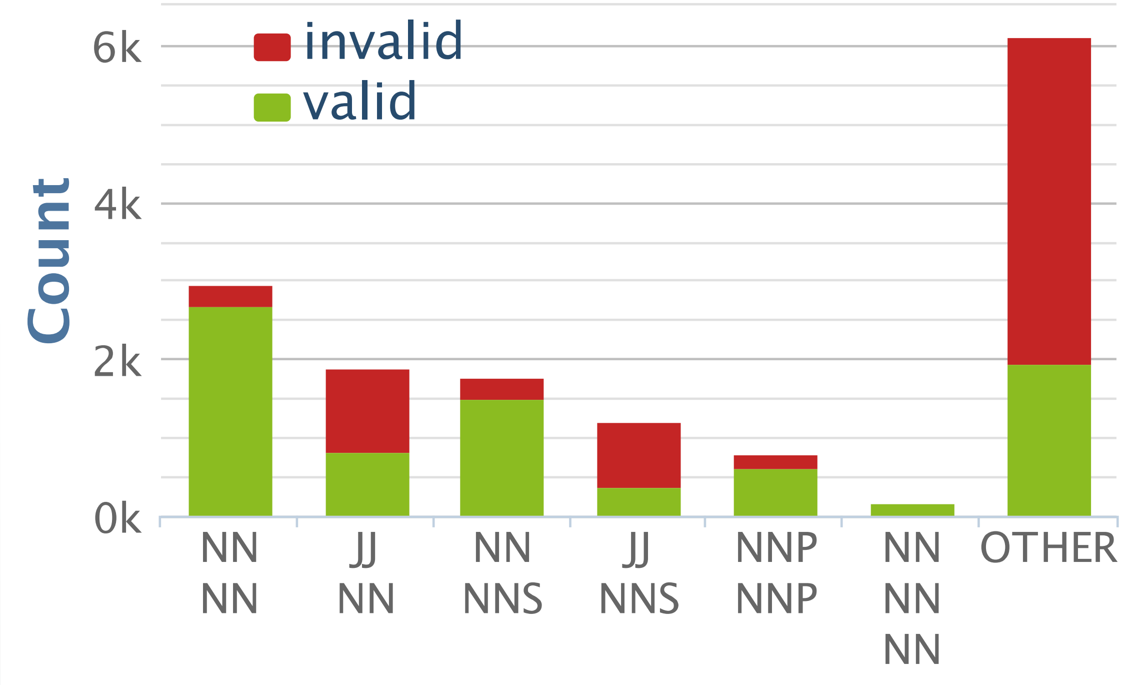

Mention
Coref
Mention
Dubrovnik is a city located in the region of Dalmatia.
Coref
Mention
This city hosted the Summer Festival in August 2016.

Named Entity Extraction: Features

- Words:
  - Words in window before and after mention
  - Sequences (n-grams), frequencies
  - Bags of words
- Morphology:
  - Capitalisation: is upper case (*China*), all upper case (*IBM*), mixed case (*eBay*)
  - Symbols: contains $, £, €, roman symbols (*IV*), ..
  - Contains period (*google.com*), apostrophe (*Mandy's*), hyphen (*speed-o-meter*), ampersand (*Fisher & Sons*)
  - Stem or Lemma (*cats -> cat*), prefix (*disadvantages -> dis*), suffix (*cats -> s*), interfix (*speed-o-meter -> o*)

Named Entity Extraction: Features

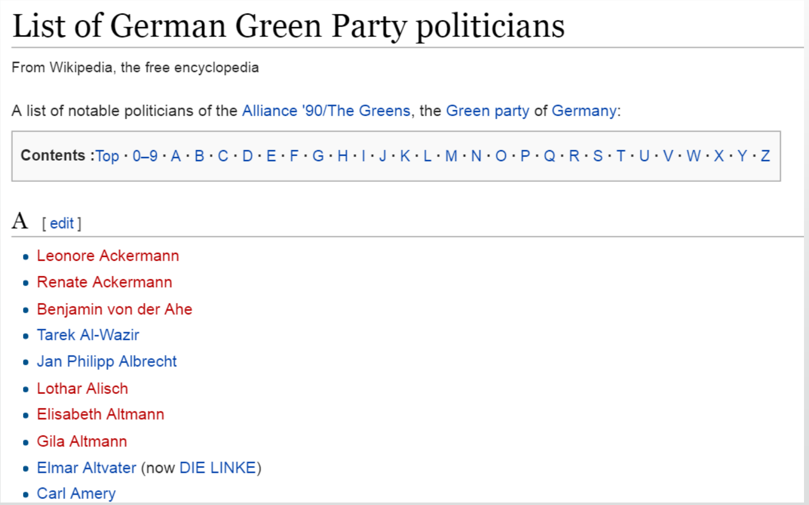
- POS tags, POS tag patterns
  - NN and NNS singular and plural nouns
  - NNP proper nouns
  - JJ adjectives



•Prokofyev (2014)

Named Entity Extraction: Features

- Gazetteers
  - Using regular expressions patterns and search engines (e.g. “*Popular artists such as * *”)
  - Retrieved from knowledge bases
    - General: Wikipedia, DBpedia, Freebase, Wikidata
    - Domain-specific: DBLP, Physics Concepts, etc.
  - Retrieved from the Web tables and lists



List of German Green Party politicians

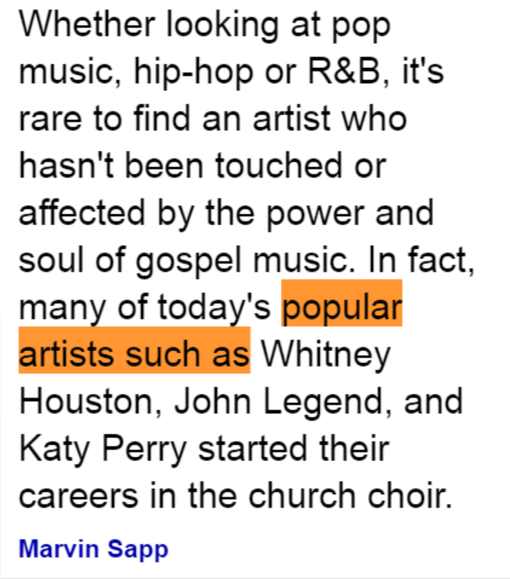
From Wikipedia, the free encyclopedia

A list of notable politicians of the Alliance '90/The Greens, the Green party of Germany:

Contents :Top · 0–9 · A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · X · Y · Z

A [edit]

- Leonore Ackermann
- Renate Ackermann
- Benjamin von der Ahe
- Tarek Al-Wazir
- Jan Philipp Albrecht
- Lothar Alisch
- Elisabeth Altmann
- Gila Altmann
- Elmar Altvater (now DIE LINKE)
- Carl Amery



Whether looking at pop music, hip-hop or R&B, it's rare to find an artist who hasn't been touched or affected by the power and soul of gospel music. In fact, many of today's popular artists such as Whitney Houston, John Legend, and Katy Perry started their careers in the church choir.

Marvin Sapp

https://en.wikipedia.org/wiki/List_of_German_Green_Party_politicians
http://www.brainyquote.com/quotes/keywords/artist.html

Named Entity Extraction: Evaluation Measures

- Precision
  - Proportion of correctly extracted NEs among all extracted NEs.
- Recall
  - Proportion of NEs found by the algorithm to all NEs in the collection.
- F-Measure
  - The weighted harmonic mean of precision and recall.

Open Problems in NER

- Extraction does not work equally well in all domains
  - Specialised technical texts
  - Other languages / multilingual text collections
- Newly emerging / unknown entities (e.g. in the context of news events)
  - Edward Snowden before the NSA scandal
  - Local entities (e.g. not widely known politicians)
  - Annotating named entities in local news papers
- Entity evolution (entity name or attribute changes over time)
  - St. Petersburg vs. Leningrad and Petrograd
  - Pope Francis vs. Jorge Mario Bergoglio

"…Bombay, also known as **Mumbai**…

-January 09, 2000 - Arts – Article
NYTimes

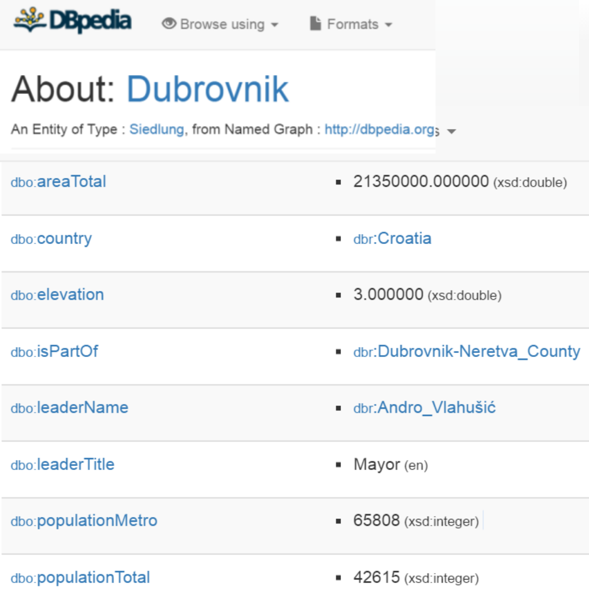
Entity Linking

- Entity Linking (EL): detecting entities and linking them to the entries of a Knowledge Base

Dubrovnik is located in the region of Dalmatia.

-> http://dbpedia.org/resource/Dubrovnik

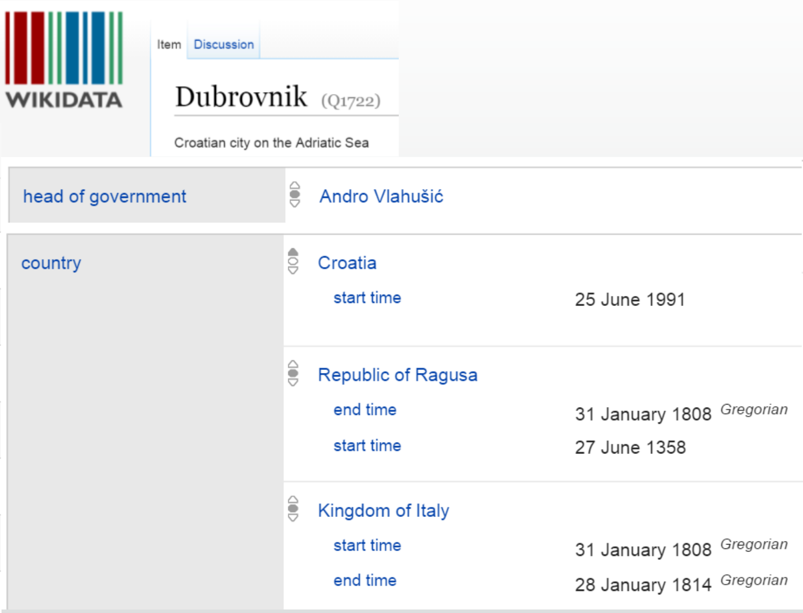
-> https://www.wikidata.org/wiki/Q1722



About: Dubrovnik

An Entity of Type : Siedlung, from Named Graph : http://dbpedia.org

dbo:areaTotal	▪ 21350000.000000 (xsd:double)
dbo:country	▪ dbr:Croatia
dbo:elevation	▪ 3.000000 (xsd:double)
dbo:isPartOf	▪ dbr:Dubrovnik-Neretva_County
dbo:leaderName	▪ dbr:Andro_Vlahušić
dbo:leaderTitle	▪ Mayor (en)
dbo:populationMetro	▪ 65808 (xsd:integer)
dbo:populationTotal	▪ 42615 (xsd:integer)



Dubrovnik (Q1722)

Croatian city on the Adriatic Sea

head of government	Andro Vlahušić		
country	Croatia		
	start time	25 June 1991	
	Republic of Ragusa		
	end time	31 January 1808	Gregorian
	start time	27 June 1358	
	Kingdom of Italy		
	start time	31 January 1808	Gregorian
	end time	28 January 1814	Gregorian

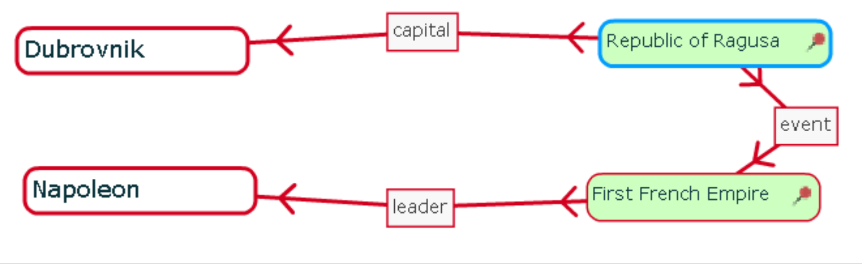
Entity Linking: Motivation

- Provide additional information / facts about the entities in the text
- Uncover relations between entities

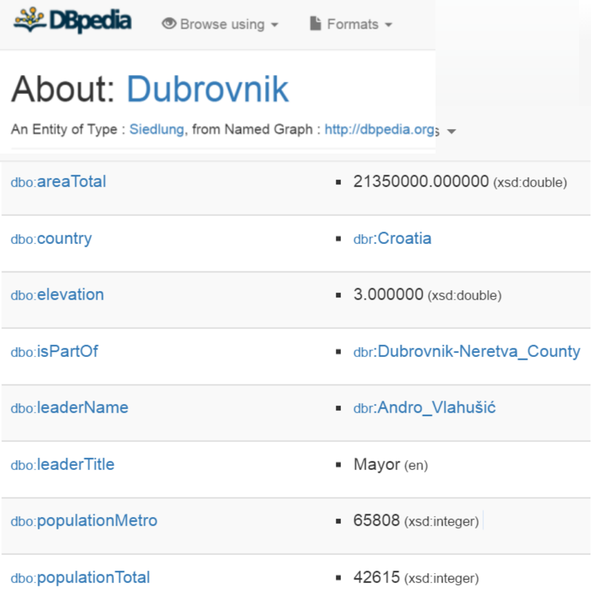
Dubrovnik

-> http://dbpedia.org/resource/Dubrovnik

-> https://www.wikidata.org/wiki/Q1722



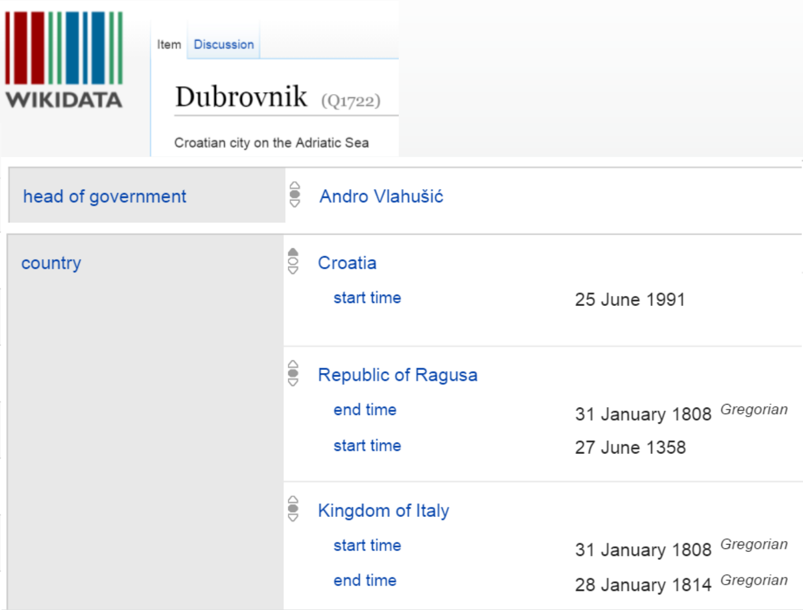
http://www.visualdataweb.org/relfinder/relfinder.php



About: Dubrovnik

An Entity of Type : Siedlung, from Named Graph : http://dbpedia.org

dbo:areaTotal	21350000.000000 (xsd:double)
dbo:country	dbr:Croatia
dbo:elevation	3.000000 (xsd:double)
dbo:isPartOf	dbr:Dubrovnik-Neretva_County
dbo:leaderName	dbr:Andro_Vlahušić
dbo:leaderTitle	Mayor (en)
dbo:populationMetro	65808 (xsd:integer)
dbo:populationTotal	42615 (xsd:integer)



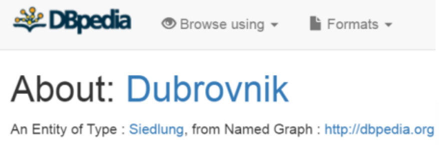
Dubrovnik (Q1722)

Croatian city on the Adriatic Sea

head of government	Andro Vlahušić	
country	Croatia	
	start time	25 June 1991
	Republic of Ragusa	
	end time	31 January 1808 Gregorian
	start time	27 June 1358
	Kingdom of Italy	
	start time	31 January 1808 Gregorian
	end time	28 January 1814 Gregorian

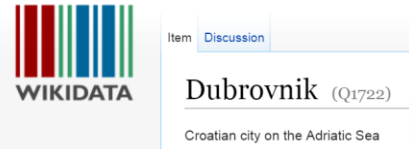
Entity Linking: Related Problems

- **KB Population**
  - Populate a KB with named entities identified in text



About: Dubrovnik

An Entity of Type : Siedlung, from Named Graph : http://dbpedia.org



Dubrovnik (Q1722)

Croatian city on the Adriatic Sea

- **Merging multiple databases**
  - Determine records represent the same entity to be merged (referred to as object identification, data de-duplication, entity resolution, entity disambiguation and record linkage)

- **Co-reference resolution or entity resolution**
  - Clustering entity mentions either within the same document or across multiple documents together, where each cluster corresponds to a single real-world entity

Hong-jie Dai , Chi-yang Wu , Richard Tzong-han , Tsai Wen-lian Hsu
From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques

Entity Linking: Pipeline

- **Spotting**
  - Detecting all non-overlapping strings in a text that could mention an entity
  - Methods: Named Entity Recognition, detecting multi word entities, finding sequences of capitalized words, surface form dictionary
- **Candidate generation**
  - Finding all possible candidate entities in KB that may be referred to the spotted string
  - Methods: query expansion and matching
- **Candidate disambiguation**
  - Selection of the most likely candidate in KB (if any)
  - Methods: ranking, classification

Entity Linking: Disambiguation Challenges

- **Name variation / evolution**
  - The same entity can be referred to by different names
    - Pope Francis, Franciscus, Jorge Mario Bergoglio
  - Methods: Dictionary
- **String ambiguity**
  - The same name string can refer to more than one entity
    - Eclipse IDE vs. Eclipse film
  - Methods: Use of context
- **Absence**
  - Many mentioned entities may not appear in a KB (NIL)
  - Methods: Classification, thresholds

Entity Linking: Disambiguation Features

- **Statistics**
  - TF-IDF (frequency and inverse frequency of candidates)
- **Entity context similarity**
  - Context in the observed phrase and in the textual description in the KB
  - Dependencies among entities in text and KB
- **Entity type information**
  - Restrictions to specific types (e.g. PERSON, LOC, ORG or domain-specific)
- **Link-based measures**
  - Hyperlinks between entities in KB (normalised inlink count or PageRank)
  - Anchor text (e.g. in Wikipedia)

Entity Linking: Tools

- **DBpedia Spotlight (hands on session)**
  - DBpedia annotations http://wiki.dbpedia.org/projects/dbpedia-spotlight
- **AIDA**
  - Maps mentions of ambiguous names onto canonical entities (e.g., individual people or places) registered in the YAGO2 knowledge base https://github.com/yago-naga/aida
- **Illinois Wikifier**
  - Disambiguating concepts and entities in a context sensitive way in Wikipedia https://cogcomp.cs.illinois.edu/page/software_view/Wikifier
- **Babelfy**
  - Babelfy is a unified, multilingual, graph-based approach to Entity Linking and Word Sense Disambiguation. http://babelfy.org/about

Temporal Extraction

- **Temporal extraction** is the extraction and normalization of temporal expressions.
- **TimeML**: specification language for temporal expressions (Pustejovsky et al. 2005)
- **Types of temporal expressions in TimeML**
  - Date: “September 7, 2016”, “tomorrow”.
  - Time: “11 a.m.”, “3 in the afternoon”.
  - Duration (length of an interval): “three years”, “since yesterday”.
  - Set (periodical aspect of an event): “twice a month”.

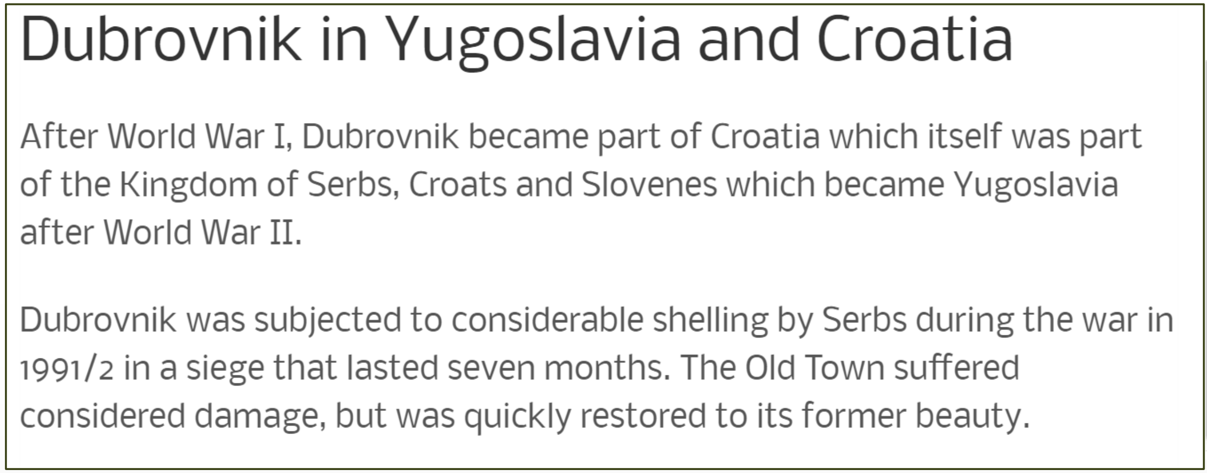
Temporal Extraction: Pipeline

- **Input**
  - Linguistic pre-processing documents with sentence, token, and POS tagging. A publication date (for news).
- **Pattern extraction**
  - Methods: Rule-based or as a classification problem: decide whether a token is part of a temporal expression.
  - Features, patterns: Terms frequently used to form temporal expressions; Names of months and weekdays or numbers that may refer to a day or year; POS, context information.
- **Normalization** (Assign temporal expression a value in a standard format)
  - Methods: Rule-based. E.g. “January” -> “01”.
- **Output**: Type (date, time, duration, set) and value

Temporal Extraction

- **Explicit expressions**
  - Fully specified and can thus be normalized without any further knowledge
  - September 6, 2016
- **Implicit expressions**
  - Names of days and events that can directly be associated with a point or interval in time
  - ESWC Summer School 2016, Christmas 2016, World War II.
- **Relative expressions**
  - Require context to normalize
  - “Today”, “the following year”

Temporal Extraction: Examples



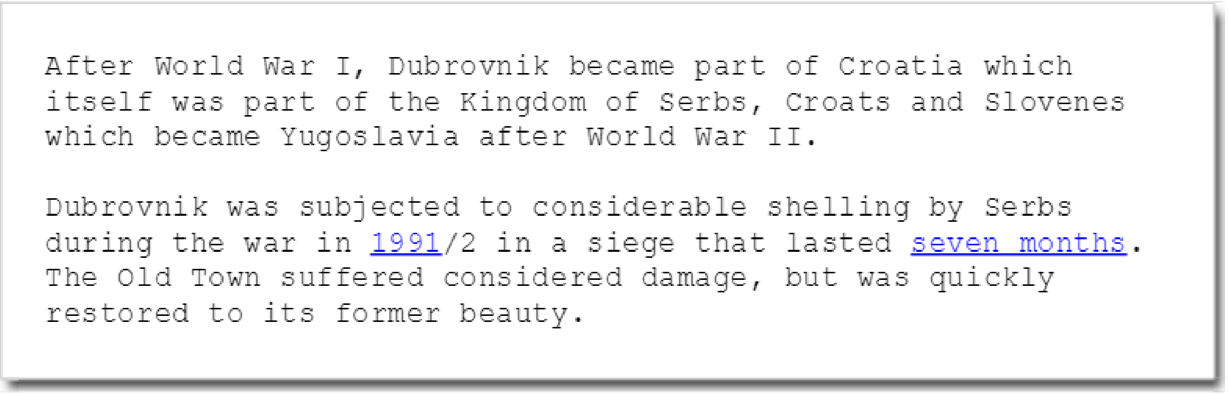
Dubrovnik in Yugoslavia and Croatia

After World War I, Dubrovnik became part of Croatia which itself was part of the Kingdom of Serbs, Croats and Slovenes which became Yugoslavia after World War II.

Dubrovnik was subjected to considerable shelling by Serbs during the war in 1991/2 in a siege that lasted seven months. The Old Town suffered considered damage, but was quickly restored to its former beauty.

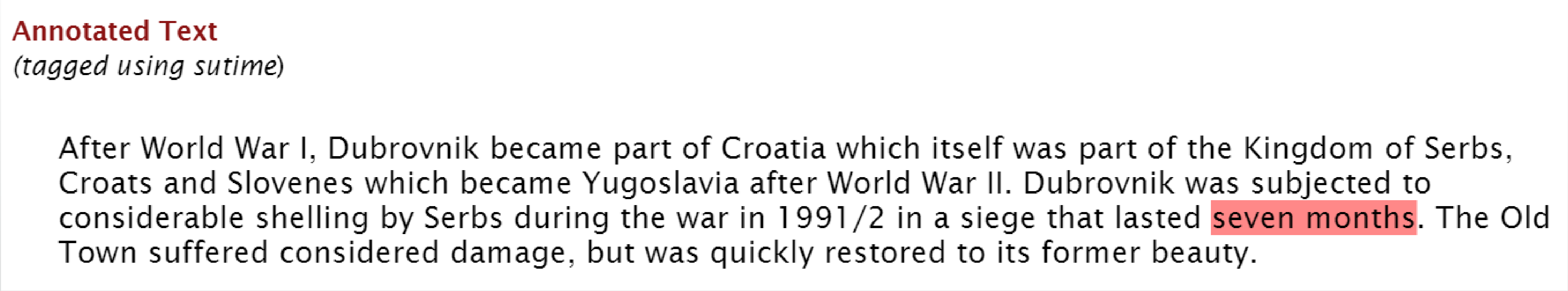
Tagged using Heideltime

Resulting document:



```
After World War I, Dubrovnik became part of Croatia which
itself was part of the Kingdom of Serbs, Croats and Slovenes
which became Yugoslavia after World War II.

Dubrovnik was subjected to considerable shelling by Serbs
during the war in 1991/2 in a siege that lasted seven months.
The Old Town suffered considered damage, but was quickly
restored to its former beauty.
```



Annotated Text

(tagged using sutime)

After World War I, Dubrovnik became part of Croatia which itself was part of the Kingdom of Serbs, Croats and Slovenes which became Yugoslavia after World War II. Dubrovnik was subjected to considerable shelling by Serbs during the war in 1991/2 in a siege that lasted seven months. The Old Town suffered considered damage, but was quickly restored to its former beauty.

Any issues?

http://www.visit-croatia.co.uk/index.php/croatia-destinations/dubrovnik/history-dubrovnik/

Temporal Extraction: Tools

- **HeidelTime** (Strötgen 2015)
  - Online demo: http://heideltime.ifi.uni-heidelberg.de/heideltime/
- **SUTime** (Angel 2012)
  - Online demo: http://nlp.stanford.edu:8080/sutime/process

Relation Extraction

- **Relations**: Two or more entities, which relate to one another in real life.
  - A relation is a tuple t = (e_1, e_2, …, e_n) where the e_i are entities in a predefined relation *R* within document *D*.
- **Binary relations**: A relation between two entities.
  - located-in(Dubrovnik, Croatia), married-to(Angelina Jolie, Brad Pitt).
- **Higher-order relations**:
  - "At codons 12, the occurrence of point mutations from G to T were observed" a 4-ary biomedical relation, a type of variation, its location, and the corresponding state change from an initial-state to an altered-state can be extracted as:
  - point mutation(codon, 12, G, T).
- **Relation extraction**: Detecting relations between entities and assigning relation types to them.

Relation Extraction: Features

- Syntactic features
  - the entities
  - the types of the entities
  - word sequence between the entities
  - number of words between the entities
- Semantic features
  - the path between the two entities in the dependency parse

Relation Extraction: Methods

- Supervised learning
  - E.g. as binary classification
- Semi-supervised and bootstrapping approaches
- Unsupervised
- Deep learning

Relation Extraction: Supervised Methods

- Relation extraction as a binary classification problem
  - Given a set of features extracted from the sentence *S*, decide if entities in *S* are connected using given relation *R*.
- Disadvantages of supervised methods
  - Need for labelled data. Difficult to extend to new relation types.
  - Extensions to higher order entity relations are difficult as well.
  - Do not scale well with increasing amounts of input data.
  - Errors in the pre-processing (feature extraction, e.g. parse tree) affect the performance.
  - Pre-defined relations only.

Relation Extraction: Semi-supervised Methods

- Semi-supervised and bootstrapping approaches (e.g. KnowItAll (Etzioni et al., 2005) and TextRunner(Banko et al., 2007))
  - Require a small set of tagged seed instances or a few hand-crafted extraction patterns per relation to launch the training process.
  - Use the output of the weak learners as training data for the next iteration.

- **Step1**: Use the seed examples to label some data.
- **Step2**: Induce patterns from the labelled examples.
- **Step3**: Apply the patterns to data, to get a new set of pairs.
- **Return to Step2**, and **iterate** until convergence criteria is reached.

Relation Extraction: Semi-supervised Methods

- Relation to be extracted **(author, book)**
- **Step1**: Use the seed examples to label data.
  - Start with one seed (Arthur Conan Doyle, The Adventures of Sherlock Holmes).
  - Pattern [order, author, book, prefix , suffix , middle].
  - Order = 1 if the author string occurs before the book string and 0 otherwise
  - Prefix and suffix are strings of 10 characters to the left/right of the match
  - Middle is the string occurring between the author and book.

Examples from DIPRE (Brin, 1998)

Relation Extraction: Semi-supervised Methods

- **Step1 (continued)**: Use the seed examples to label data.
- [order, author, book, prefix , suffix , middle].
- (Arthur Conan Doyle, The Adventures of Sherlock Holmes)
- *S1*="know that Sir Arthur Conan Doyle *wrote* The Adventures of Sherlock Holmes, in 1892"

Relation Extraction: Semi-supervised Methods

- **Step1 (continued)**: Use the seed examples to label data.
- [order, author, book, prefix , suffix , middle].
- (Arthur Conan Doyle, The Adventures of Sherlock Holmes)
- *S1*="know that Sir Arthur Conan Doyle *wrote* The Adventures of Sherlock Holmes, in 1892"
- [1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, know that Sir, in 1892, *wrote*]

Relation Extraction: Semi-supervised Methods

- **Step1 (continued)**: Use the seed examples to label data.
- [order, author, book, prefix , suffix , middle].
- (Arthur Conan Doyle, The Adventures of Sherlock Holmes)
- *S1*=“know that Sir Arthur Conan Doyle *wrote* The Adventures of Sherlock Holmes, in 1892”
- [1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, know that Sir, in 1892, *wrote*]
- *S2*=“When Sir Arthur Conan Doyle *wrote* the adventures of Sherlock Holmes in 1892 he was high …”
- [1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, When Sir, in 1892 he, *wrote*]

Relation Extraction: Semi-supervised Methods

- **Output Step1:** [order, author, book, prefix , suffix , middle].

  [1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, now that Sir, in 1892, wrote]

  [1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, When Sir, in 1892 he, wrote]

- **Step2:** Induce patterns from the labelled examples.
  - Exact match: Generalize the pattern: [Sir, .*?, *wrote*, .*?, in 1892].
  - Approximate match: Use similarity metrics for patterns. Agichtein & Gravano, 2000)
- **Step3:** Apply the patterns to data, to get a new set of pairs.
  - (Arthur Conan Doyle, The Speckled Band).
- **Return to Step2**, and **iterate** until convergence criteria is reached.

Open Information Extraction (Open IE)

- Open IE extracts tuples consisting of argument phrases and a relation between the arguments
  - (arg1; rel; arg2).
- For example: *S3*="Romney (arg1) will be elected (pred) President (arg2)."
  - (Romney; will be elected; President)
- Different to relation extraction
  - No pre-specified sets of relations
  - No domain-specific knowledge engineering
- Example applications
  - A news reader who wishes to keep abreast of important events
  - An intelligence analyst who recently acquired a terrorist's laptop

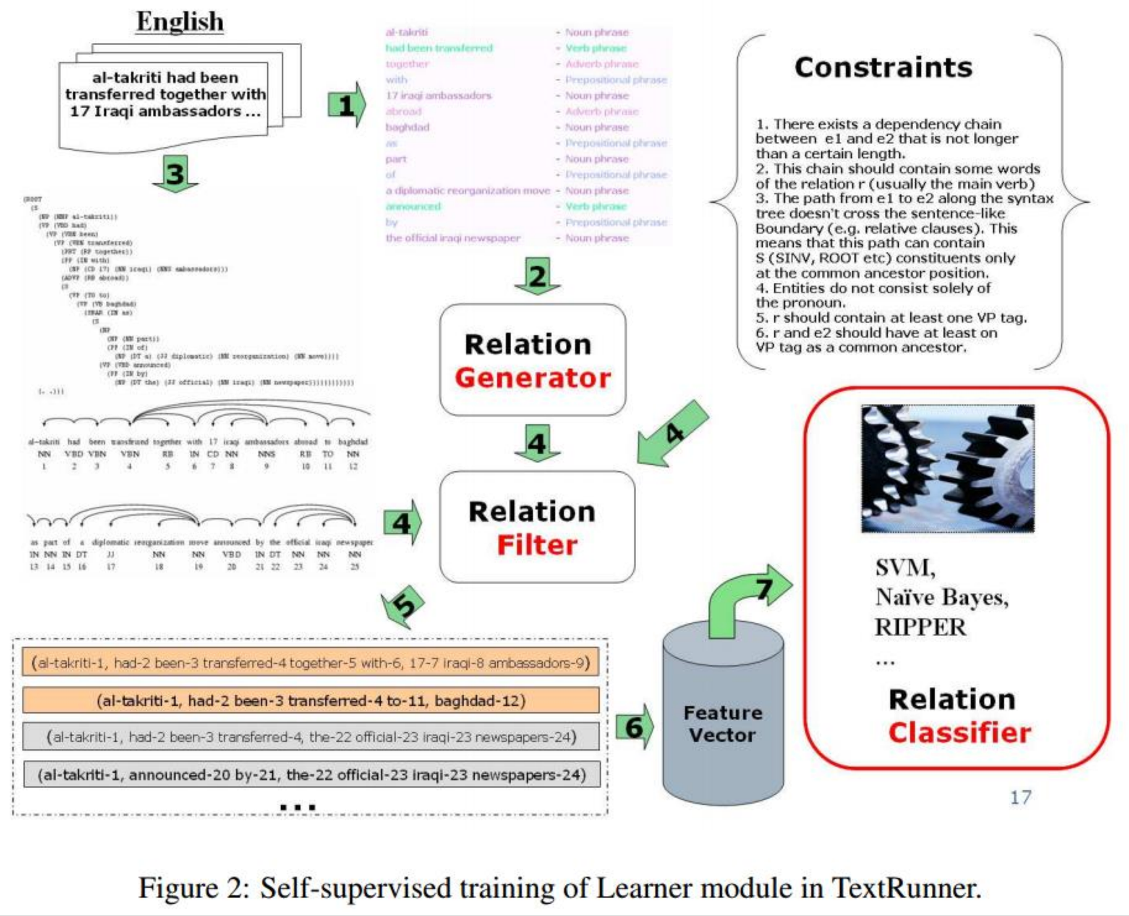
Open IE: TextRunner

(Banko et al., 2007)

- **Relation is a tuple:** $t = (e_1, r, e_2)$
  - e_1 and e_2 are surface forms of entities or noun phrases.
  - r denotes a relationship between e_1 and e_2.
- **TextRunner components:**
  - Self-supervised Learner: automatic labelling of training data.
  - Single-pass Extractor: generates candidate relations from each sentence, runs a classifier and retains the ones labelled as trustworthy relations.
  - Redundancy-based Assesor: assigns a probability to each retained tuple based on a probabilistic model of redundancy in text.

Electronics and
Computer Science

Open IE: TextRunner


English
al-takriti had been transferred together with 17 Iraqi ambassadors ...
1
al-takriti - Noun phrase
had been transferred - Verb phrase
together - Adverb phrase
with - Prepositional phrase
17 iraqi ambassadors - Noun phrase
abroad - Adverb phrase
baghdad - Noun phrase
as - Prepositional phrase
part - Noun phrase
of - Prepositional phrase
a diplomatic reorganization move - Noun phrase
announced - Verb phrase
by - Prepositional phrase
the official iraqi newspaper - Noun phrase
2
3
Relation Generator
4
4
4
Relation Filter
5
6
Feature Vector
7
SVM,
Naïve Bayes,
RIPPER
...
Relation Classifier
17

Constraints

1. There exists a dependency chain between e1 and e2 that is not longer than a certain length.
2. This chain should contain some words of the relation r (usually the main verb)
3. The path from e1 to e2 along the syntax tree doesn't cross the sentence-like Boundary (e.g. relative clauses). This means that this path can contain S (SINV, ROOT etc) constituents only at the common ancestor position.
4. Entities do not consist solely of the pronoun.
5. r should contain at least one VP tag.
6. r and e2 should have at least on VP tag as a common ancestor.

(al-takriti-1, had-2 been-3 transferred-4 together-5 with-6, 17-7 iraqi-8 ambassadors-9)

(al-takriti-1, had-2 been-3 transferred-4 to-11, baghdad-12)

(al-takriti-1, had-2 been-3 transferred-4, the-22 official-23 iraqi-23 newspapers-24)

(al-takriti-1, announced-20 by-21, the-22 official-23 iraqi-23 newspapers-24)

...

Figure 2: Self-supervised training of Learner module in TextRunner.

(Banko et al., 2007) image: (Bach 2013)

- **7-step training** for each sentence
- **Step 1**: a noun phrase chunker.
- **Step 2**: the relation candidator.
- **Steps 3-5**: a syntactic parser and dependency parser are run. The relation filter uses parse trees, dependency trees, and set of constraints to label trustworthy and untrustworthy relations.
- **Step 6**: map each relation to a feature vector representation.
- **Step 7**: train a binary classifier using labelled trustworthy and untrustworthy relations.

Open IE: REVERB and Ollie

- REVERB (Fader 2011) and Ollie (Mausam 2012) extract binary relationships from English sentences.
- Designed for Web-scale information extraction, where the target relations cannot be specified in advance and speed is important.
- REVERB extracts relations mediated by verbs, does not consider the context
  - shallow syntactic processing to identify relation phrases that begin with a verb and occur between the argument phrases.
- Ollie extracts relations mediated by nouns, adjectives, and more. Ollie includes contextual information from the sentence in the extractions.
- http://reverb.cs.washington.edu/ https://knowitall.github.io/ollie/

Open IE: Ollie System Architecture

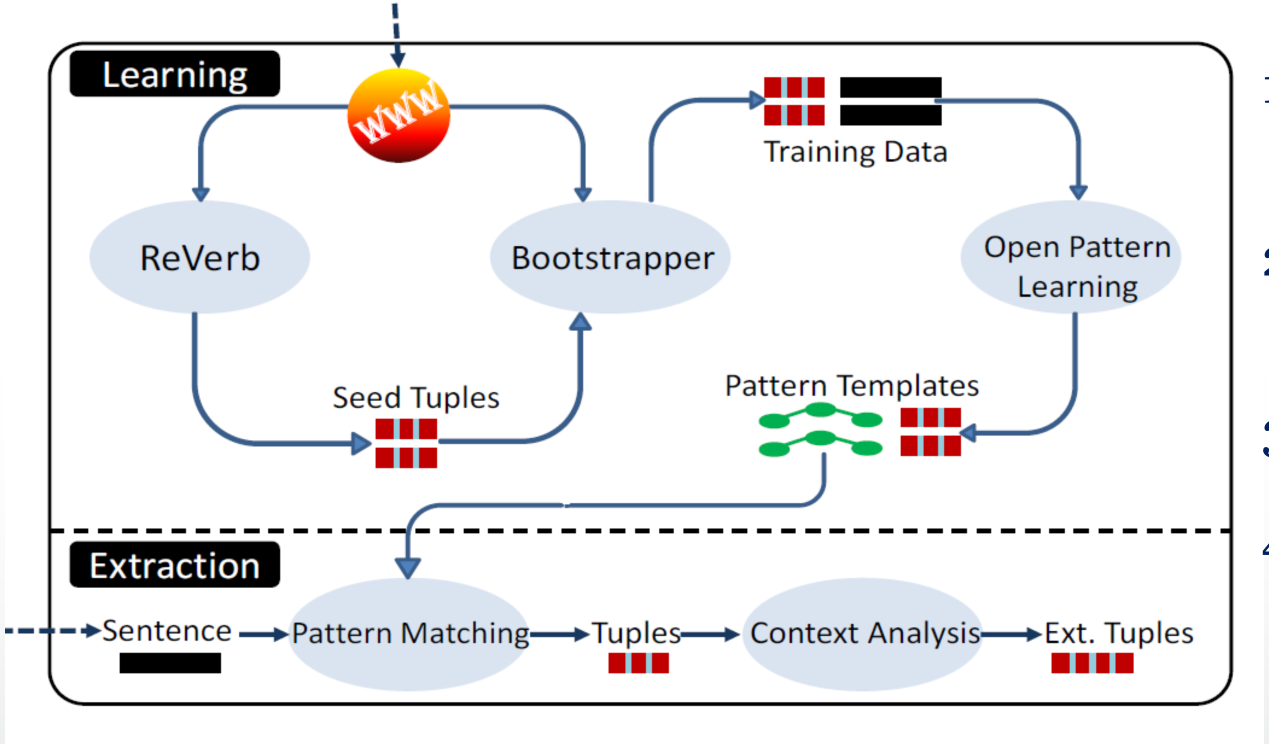


Image from (Mausam 2012)

1) Use a set tuples from REVERB to bootstrap a large training set.

2) Learn open pattern templates over this training set.

3) Apply pattern templates at extraction time.

4) Analyse the context around the tuple to add information (attribution, clausal modifiers) and a confidence function.

Open IE: Ollie Bootstrapping

- Goal is to automatically create a large training set, which encapsulates the multitudes of ways in which information is expressed in text.
- Almost every relation can also be expressed via a REVERB-style verb-based expression.
- Retrieve all sentences in a Web corpus that contains all content words in the tuple.
- Assumption: sentences express the information of the original seed tuple. Not always true:
  - (Boyle; is born in; Ireland)
  - “Felix G. Wharton was born in Donegal, in the northwest of Ireland, a county where the Boyles did their schooling.”

Open IE: Ollie Bootstrapping

- Over 110,000 seed tuples –high confidence REVERB extractions from ClueWeb - a large Web corpus. Contain only proper nouns in the arguments.
  - Seed: “Paul Annacone is the coach of Federer.” ->
  - REVERB pattern: (Paul Annacone; is the coach of; Federer).
  - Retrieved sentence: “Now coached by Annacone, Federer is winning more titles than ever.”
- Enforce additional dependency restrictions on the sentences to reduce bootstrapping errors.
- Restrict linear path length between argument and relation in the dependency parse (max 4).

REVERB vs. Ollie

- “Early astronomers believed that the earth is the center of the universe.”
  - R: (the earth; be the center of; the universe)
  - O: ((the earth; be the center of; the universe) AttributedTo believe; Early astronomers)
- “If he wins five key states, Romney will be elected President.”
  - R: (Romney; will be elected; President)
  - O: ((Romney; will be elected; President) ClausalModifier if; he wins five key states)

NLP & ML Software

Natural Language Processing:

- Stanford NLP (Java)
- GATE (general purpose architecture, includes other NLP and ML software as plugins)
- OpenNLP (Java)
- NLTK (Python)

Machine Learning:

- scikit-learn (Python)
- Mallet (Java)
- WEKA (Java)
- Alchemy (graphical models, Java)
- FACTORIE, wolfe (graphical models, Scala)
- CRFSuite (efficient implementation of CRFs, Python)

NLP & ML Software

Ready to use NERC software:

- ANNIE (rule-based, part of GATE)
- Wikifier (based on Wikipedia)
- FIGER (based on Wikipedia, fine-grained Freebase NE classes)

Almost ready to use NERC software:

- CRFSuite (already includes Python implementation for feature extraction, you just need to feed it with training data, which you can also download)

Ready to use RE software:

- ReVerb, Ollie (Open IE, extract patterns for any kind of relation)
- MultiR (Distant supervision, relation extractor trained on Freebase)

Web content extraction software:

- Boilerpipe (extract main text content from Web pages)
- Jsoup (traverse elements of Web pages individually, also allows to extract text)

Thank you!

Questions, Comments?

Dr. Elena Demidova

University of Southampton
Web and Internet Science Group
email: e.demidova@soton.ac.uk
www: https://demidova.wordpress.com

This presentation contains slides by Isabelle Augenstein, "Information Extraction with Linked Data Tutorial", ESWC Summer School 2015.

References

Entity Linking

- H. Dai, C. Wu, R. Tsai, and W. Hsu. From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques. In The 26th Annual Conference of the Japanese Society for Artificial Intelligence, pp 1–10, 2012.

Temporal extraction

- Pustejovsky, J., Knippen, R., Littman, J., & Sauri, R. (2005). Temporal and event information in natural language text. Language resources and evaluation, 39(2–3), 23–164. 2005.
- Jannik Strötgen, Michael Gertz. Multilingual and cross-domain temporal tagging. Language Resources and Evaluation. June 2013, Volume 47, Issue 2, pp 269-298.
- Jannik Strötgen, Michael Gertz. A Baseline Temporal Tagger for All Languages. EMNLP'15.
- Angel X. Chang and Christopher D. Manning. 2012. SUTIME: A Library for Recognizing and Normalizing Time Expressions. 8th International Conference on Language Resources and Evaluation (LREC 2012).

References

- **Relation extraction**

  - N Bach, S Badaskar. A review of relation extraction. 2013
  - Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. Proceedings of IJCAI ’07.
  - Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. Proceedings of the Fifth ACM International Conference on Digital Libraries.
  - Brin, S. (1998). Extracting patterns and relations from the world wide web. WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT ’98.
  - Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. EMNLP-CoNLL '12. pp. 523-534.
  - Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In EMNLP '11. pp. 1535-1545.