Information Extraction

Tutor: Rahmad Mahendra
rahmad.mahendra@cs.ui.ac.id

Slide by: Bayu Distiawan Trisedya
Main Reference: Stanford University

Natural Language Processing & Text Mining
Short Course

Pusat Ilmu Komputer UI
22 – 26 Agustus 2016

Information Extraction

- Information extraction (IE) systems
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
  - Goals:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Extracting Information from Text

- Data stored digitally
  - Image, video, music, **text**
- What information are stored (on internet)?
- How can we use that information?

What information are stored (on internet)?

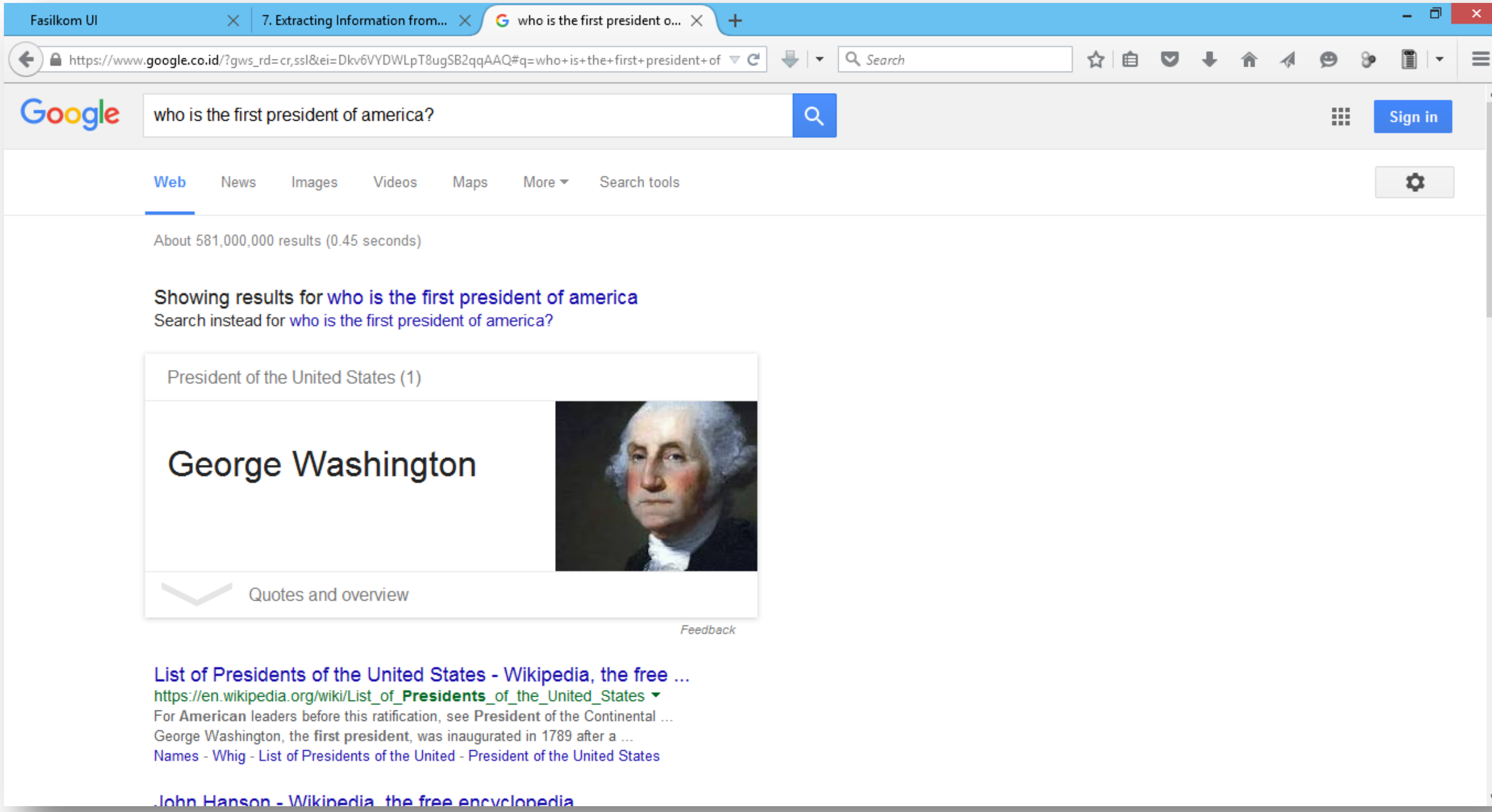
- Structured Data

Name	GPE
Barack Obama	USA
Joko Widodo	Indonesia
Malcolm Turnbull	Australia
Najib Razak	Malaysia

- Unstructured Data

“Malcolm Bligh Turnbull is the 29th and current Prime Minister of Australia and the Leader of the Liberal Party, having assumed office in September 2015. He has served as the Member of Parliament for Wentworth since 2004.”

Finding Information



Why is IE hard on the web?

A book,
Not a toy

Title

Need this price

pusilkom ui
quality results

How do we get a machine to understand the text?

- One approach to this problem:
  - Convert the unstructured data of natural language sentences into the structured data
    - Table, relational database, etc
  - Once the date are structured, we can use query tools such as SQL
- Getting meaning from text is called **Information Extraction**

pusilkom ui
quality results

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

pusilkom ui
quality results

Slide acknowledgement: Stanford NLP

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person	Date	Location	Organization

pusilkom ui
quality results

Slide acknowledgement: Stanford NLP

Three standard approaches to NER (and IE)

1. Hand-written regular expressions
2. Using classifiers
   – Naïve Bayes
3. Sequence models
   – CMMs/MEMMs

Hand-written Patterns for Information Extraction

- If extracting from automatically generated web pages, simple regex patterns usually work.
  - Amazon page
  - <div class="buying"><h1 class="parseasinTitle">(.*?)</h1>

- For certain restricted, common types of entities in unstructured text, simple regex patterns also usually work.
  - Finding phone numbers
  - (?:\(?[0-9]{3}\)?[-.])?[0-9]{3}[-.]?[0-9]{4}

pusilkom ui
quality results

Natural Language Processing-based Hand-written Information Extraction

- For unstructured human-written text, some NLP may help
  - Part-of-speech (POS) tagging
    - Mark each word as a noun, verb, preposition, etc.
  - Syntactic parsing
    - Identify phrases: NP, VP, PP
  - Semantic word categories (e.g. from WordNet)
    - KILL: kill, murder, assassinate, strangle, suffocate

Rule-based Extraction Examples

Determining which person holds what office in what organization

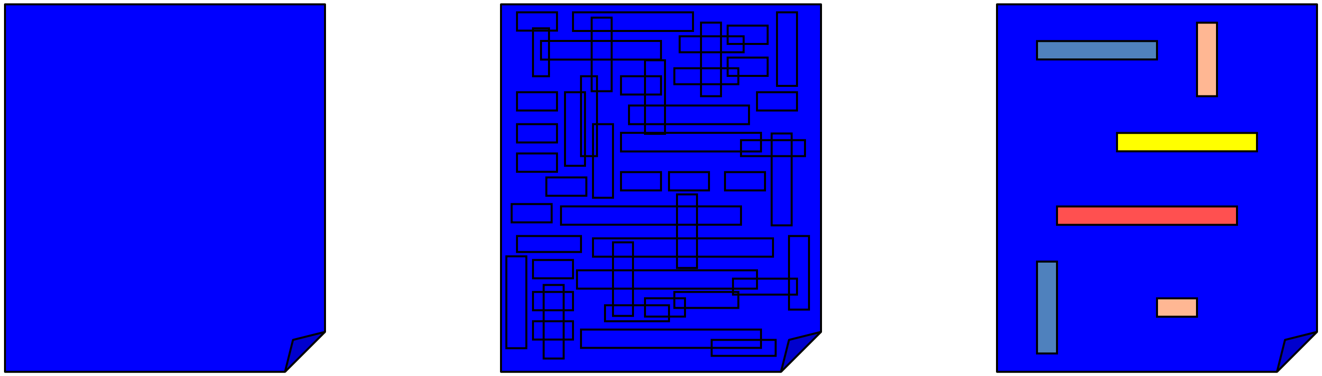
- [person] , [office] *of* [org]
  - Vuk Draskovic, leader of the Serbian Renewal Movement
- [org] (*named*, *appointed*, etc.) [person] Prep [office]
  - NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

- [org] *in* [loc]
  - NATO headquarters in Brussels
- [org] [loc] (*division*, *branch*, *headquarters*, etc.)
  - KFOR Kosovo headquarters

Naïve use of text classification for IE

- Use conventional classification algorithms to classify substrings of document as "*to be extracted*" or not.



- In some simple but compelling domains, this naive technique is remarkably effective.
  - But do think about when it would and wouldn't work!

‘Change of Address’ email

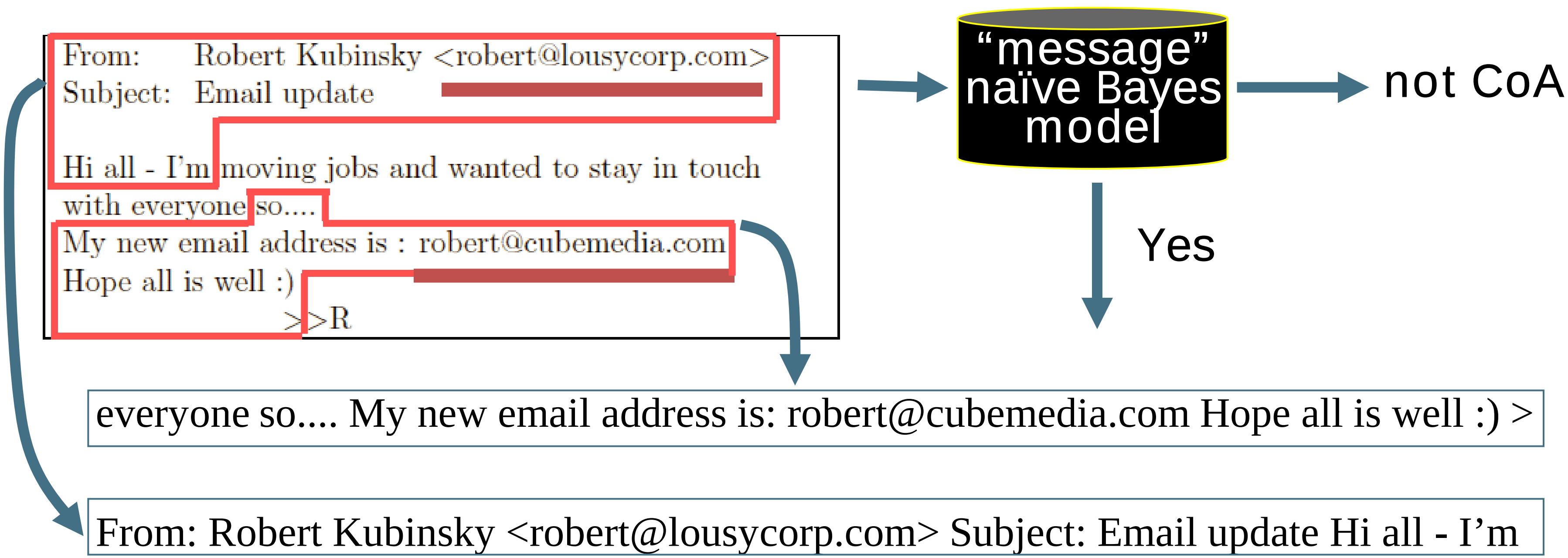
From: Robert Kubinsky <robert@lousycorp.com>
Subject: Email update

Hi all - I’m moving jobs and wanted to stay in touch with everyone so....
My new email address is : robert@cubemedia.com
Hope all is well :)
>>R

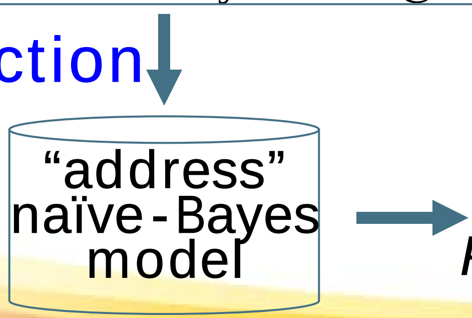
Change-of-Address detection

[Kushmerick et al., ATEM 2001]

1. Classification



2. Extraction



P[robert@lousycorp.com] = 0.28

P[robert@cubemedia.com] = 0.72

ML sequence model approach to NER

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

Slide acknowledgement: Stanford NLP

Features for sequence labeling

- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label
- Word substrings
  - Cotrim**oxa**zole, ciprofl**oxa**cin, sulfameth**oxa**zole

Slide acknowledgement: Stanford NLP

Sequence problems

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences …
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

MEMM inference in systems

- For a Conditional Markov Model (CMM) a.k.a. a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions
- A larger space of sequences is usually explored via search

Decision Point

Local Context

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
...	...

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

pusilkom ui
quality results

Slide acknowledgement: Stanford NLP

Evaluation: Precision and recall

- **Precision:** % of selected items that are correct
  Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

Evaluation Example (PER)

	Actual	Prediction
Foreign	ORG	LOC
Ministry	ORG	PER
spokesman	O	O
Shen	PER	O
Guofang	PER	PER
told	O	O
Reuters	ORG	ORG
:	:	:

	correct	not correct
selected	Tp: 1	Fp: 1
not selected	Fn: 1	

A combined measure: F

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = ½$): $F = 2PR/(P+R)$

pusilkom ui
quality results

Named Entity Recognition Task

Task: Predict entities in a text

Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

Standard evaluation is per entity, *not* per token

Precision/Recall/F1 for IE/NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings …
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics: give partial credit

Thank you

pusilkom ui

quality results