# 04 – Named Entity Recognition

## IA161 Advanced Techniques of Natural Language Processing

Z. Nevěřilová

NLP Centre, FI MU, Brno

October 8, 2015

Washington: Ben Carson said Wednesday he's pulling in lots of money amid all the backlash he's received for remarks he made regarding Muslims in politics. The retired neurosurgeon said he raised $1 million within 24 hours following the CNN debate on Sept. 16, and that donations have poured in after remarks he made over the weekend about Islam and the presidency. "The money has been coming in so fast, it's hard to even keep up with it," he said Wednesday morning on Fox News, when asked about whether his comments had affected his donations. "I remember the day of the last debate, within 24 hours we raised $1 million. And it's coming in at least at that rate if not quite a bit faster." CNN will not be able to verify fundraising totals with the Federal Election Commission until after the quarter ends Sept 30.

# Outline

# Named Entity Recognition (NER)

NER aims to recognize and classify names of people, locations, organizations, products, artworks, domain names, phone numbers, dates, money, measurements (numbers with units), law or patent numbers etc.

Named entities (NEs) can be one word or multi word.
[overlap with multi word expression (MWE) processing]

Multi word NE processing leads to retokenization.

**Example**

| | NE | MWE |
|---|---|---|
| Brno | ✓ | ✗ |
| a priori | ✗ | ✓ |
| New York | ✓ | ✓ |

# NER and information extraction (IE)

## Example

MIT Press published a book by Patrick Hanks with the title Lexical Analysis: Norms and Exploitations. [MIT Press] published a book by [Patrick Hanks] with the title [Lexical Analysis: Norms and Exploitations].

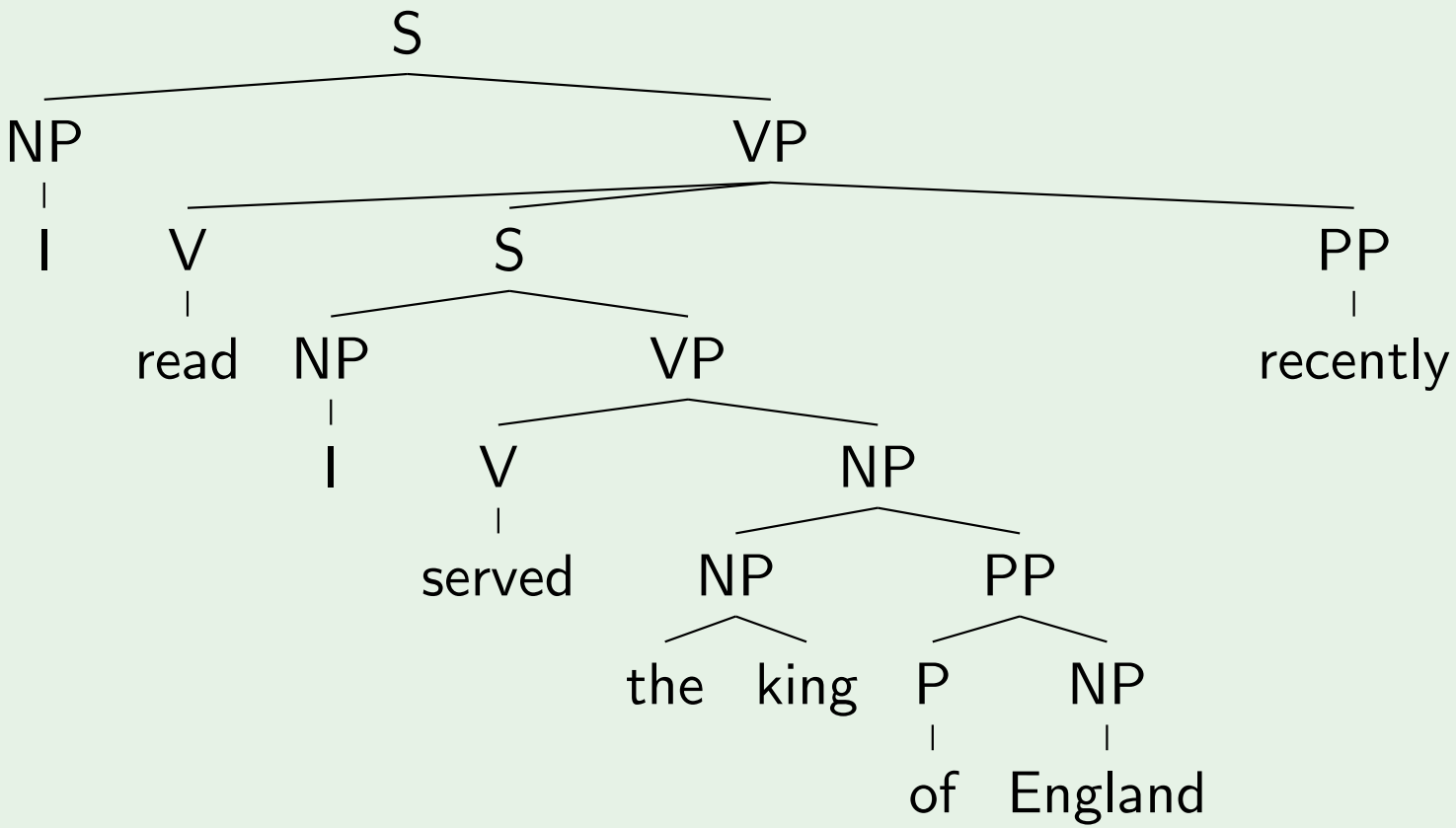
MIT Press published a book by Randy Thornhill and Craig T. Palmer entitled A Natural History of Rape: Biological Bases of Sexual Coercion [MIT Press] published a book by [Randy Thornhill] and [Craig T. Palmer] entitled [A Natural History of Rape: Biological Bases of Sexual Coercion]

| Authors | | Title |
|---|---|---|
| Patrick Hanks | | Lexical Analysis: Norms and Exploitations |
| Randy Thornhill | Craig T. Palmer | A Natural History of Rape: Biological Bases of Sexual Coercion |

# Named Entity Recognition (NER)

Retokenization can improve significantly advanced natural language processing:

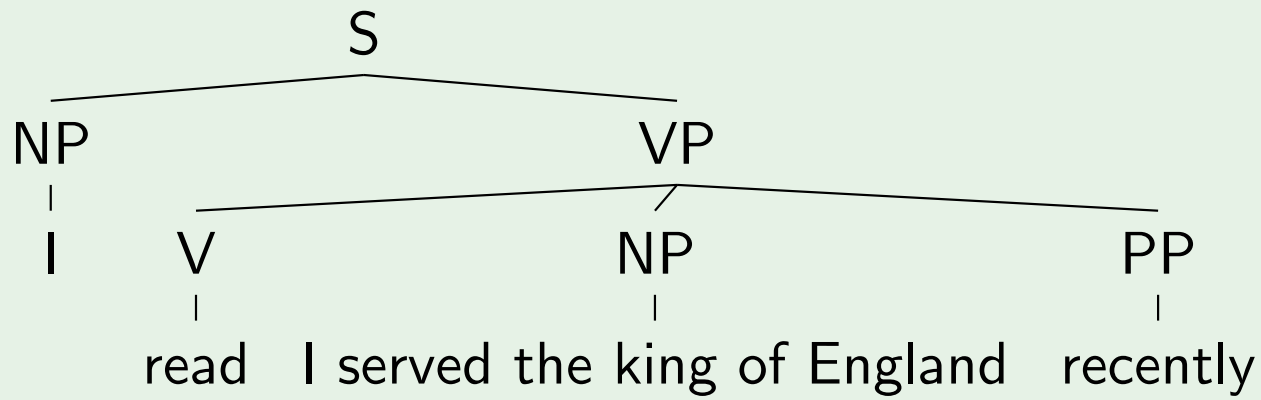
## Example



# Named Entity Recognition (NER)

Retokenization can improve significantly advanced natural language processing:

## Example



# NER: recognizing boundaries

**Example**

Masaryk University in Brno

Masaryk University in Brno

Masaryk University in Brno

**Example**

The Picture of Dorian Gray

The Picture of Dorian Gray


Nová opera Vladimíra

Franze Válka s mloky . . . Nová

# Named Entity Classification

Common classes: PERSON, ORGANIZATION, LOCATION
Less common classes: MONEY, PERCENT, DATE, TIME
Rare classes: ARTWORK, PRODUCT, ROLE

**Example**

| | |
|---|---|
| The White House | LOCATION? ORGANIZATION |
| Othello | PERSON? ARTWORK? PRODUCT? |
| Motorola | ORGANIZATION? PRODUCT? |
| The Pope | PERSON? ROLE? |
| two years ago | DATE? nothing? |

The main problem is with metonymy.

# Methods for NER

- gazetteer methods (list of NEs)
- semi-supervised machine learning (bootstraping)
- supervised machine learning (training)

# Gazetteer Methods for NER

lists of NEs + substring search algorithms:

- list of names
- list of company names
- list of place names

search all occurences of substrings $S_k, \ldots, S_l$ from lists of pattern strings $P_1, \ldots, P_p$ in a target string $T[1 \ldots m]$

- naïve multi-pass: $O(p(m+n))$
- improvements: Rabin-Karp, Boyer-Moore, Knuth-Morris-Pratt
- single-pass: Aho-Corasick: $O(m+k)$

where $p$ is the number of patterns,
$m$ is the target string length,
$n$ is the average pattern length,
$k$ is the total number of occurrences of the pattern strings in the text

# Gazetteer Methods for NER

Problems: disambiguation + fixedness

**Example**

May the force be with you!
I was born on May.
Karel May is my favorite writer.

**Example**

Google was bought by Brand New So-far-unknown Company Inc.

# Semi-supervised methods for NER

bootstrapping = a small degree of supervision
typically requires a small set of *seeds*

## Example

seeds: John, James, Steve
search patterns in contexts:
Peter, David, Michael ...

## Example

`[Capitalized words and letters]`, the CEO of
`[Capitalized words and non-capitalized stop words]`,

Richard Rosenblatt, the CEO of Demand Media,
Michael Close, the CEO of Enterprise Training Centre,
...

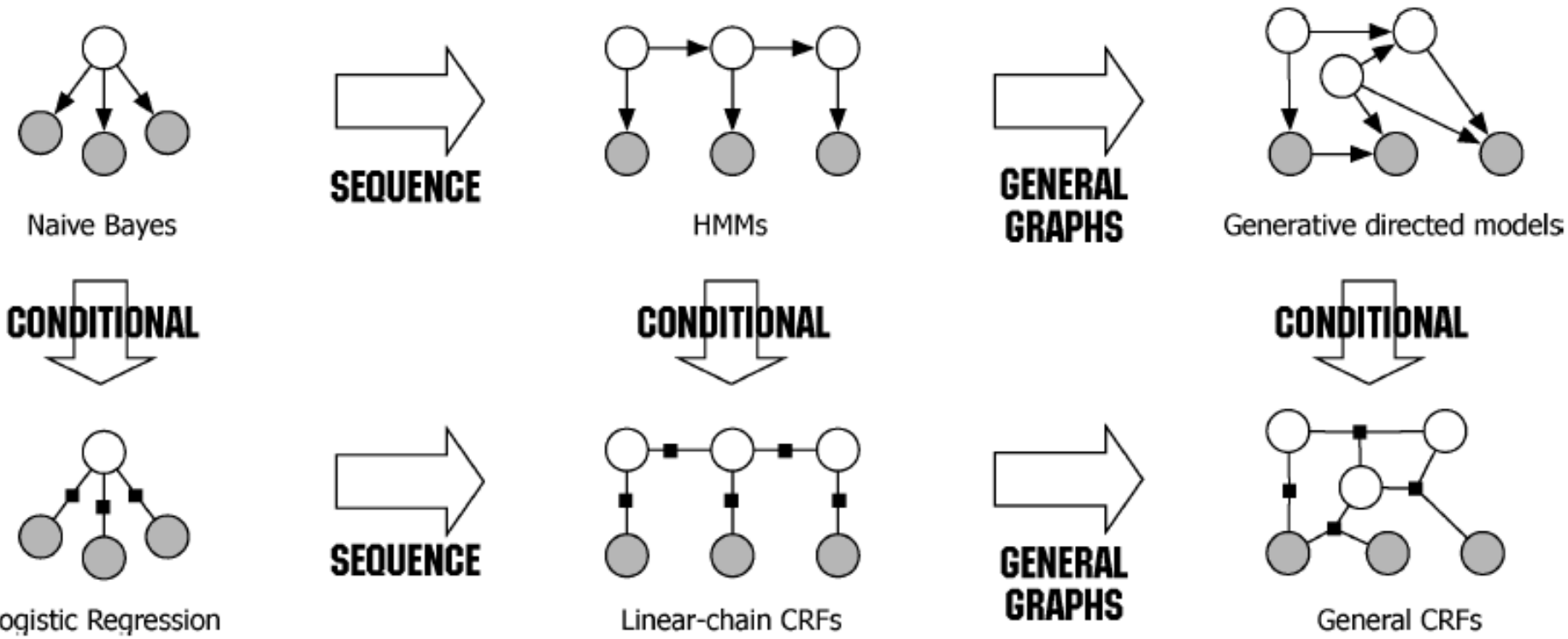
# Semi-supervised methods for NER

good for discovering NEs (fixedness problem solved)
but not good at disambiguation

# Supervised methods for NER

manually annotated training set
manually annotated test set (the golden standard)
+ optionally the gazetteer

discriminative vs. generative methods



# Evaluation of NER systems

precision, recall, F1-score
separate precision, recall, F1-score measurements for different classes
the less difficult classes are: DATE, MONEY, PERCENT
the most difficult classes are: PERSON, ORGANIZATION

Error analysis:

- errors in boundaries detection
- errors in class labeling

What is preferred: high precision (and low recall) or high recall (and more false positives)?

. . . see also [8]

# Current state-of-the-art results

| Language | System | F1 |
|---|---|---|
| English | MUC-7[1], baseline | 58.89% |
| English | MUC-7 human annotation | 97.60% |
| English | MUC-7 best result [9] | 93.39% |
| English | CONLL-2003 best result [3] | 88.76% |
| English | CONLL-2003 [6] | 90.10% |
| German | GermEval 2014 best result [5] | 77.14% |
| Russian | [4] | 75.05% |
| Czech | [11] | 82.82% |
| Czech | [7] | 83.24% |
| Arabic | [1] | 65.76% |

[1] Message Understanding Conference


Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio.
A semi-supervised learning approach to Arabic named entity recognition.
In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *RANLP*, pages 32–40. RANLP 2011 Organising Committee / ACL, 2013.


R. A. Baeza-Yates.
Algorithms for string searching.
*SIGIR Forum*, 23(3-4):34–58, April 1989.


Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang.
Named entity recognition through classifier combination.
In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.


Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov.
Introducing baselines for Russian named entity recognition.
In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'13, pages 329–342, Berlin, Heidelberg, 2013. Springer-Verlag.


Christian Hänig, Stefan Thomas, and Stefan Bordag.
Modular classifier ensemble architecture for named entity recognition on low resource systems.
2014.


Zhiheng Huang, Wei Xu, and Kai Yu.
Bidirectional LSTM-CRF models for sequence tagging.
*CoRR*, abs/1508.01991, 2015.


Michal Konkol and Miloslav Konopík.
Crf-based czech named entity recognizer and consolidation of Czech NER research.
In Ivan Habernal and Václav Matoušek, editors, *Text, Speech, and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 153–160. Springer Berlin Heidelberg, 2013.


Chris Manning.
Doing named entity recognition? Don't optimize for F1.
online, accessible on `http://nlpers.blogspot.cz/2006/08/doing-named-entity-recognition-dont.html`, accessed 2015-10-08.


Andrei Mikheev, Claire Grover, and Marc Moens.
*Description of the LTG system used for MUC-7*.
Association for Computational Linguistics, 1998.


David Nadeau and Satoshi Sekine.
A survey of named entity recognition and classification.
*Linguisticae Investigationes*, 30(1):3–26, January 2007.
Publisher: John Benjamins Publishing Company.


Jana Straková, Milan Straka, and Jan Hajič.
A new state-of-the-art Czech named entity recognizer.
In Ivan Habernal and Václav Matoušek, editors, *Text, Speech, and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 68–75. Springer Berlin Heidelberg, 2013.


Charles Sutton and Andrew McCallum.
An introduction to conditional random fields.
*Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.