

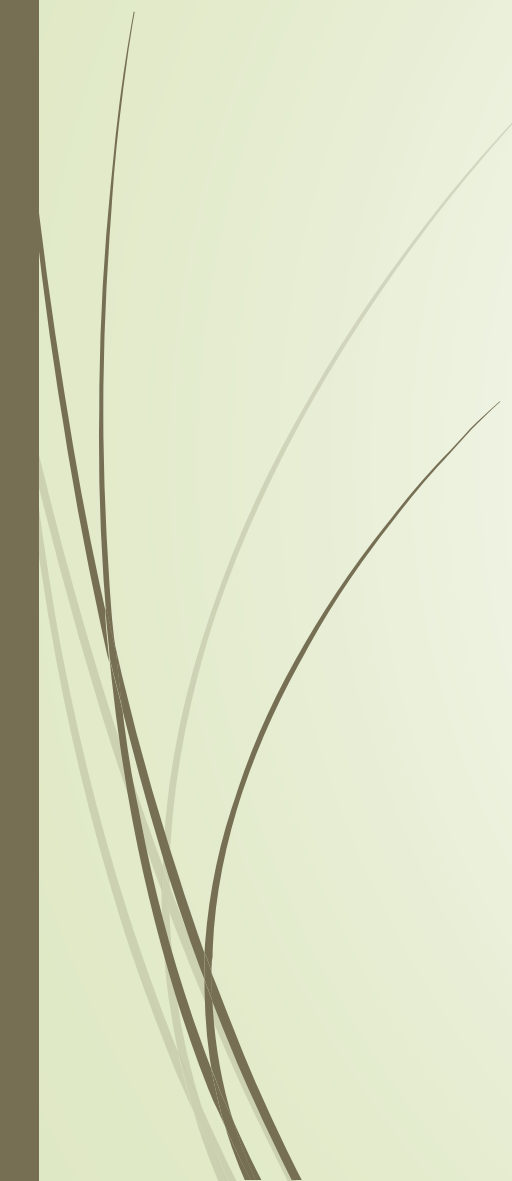


# Named Entity Recognition

Atıl Vural & Dilek Kayahan



# Outline

- Introduction
  - The approaches and methods used
  - Different Languages
  - Success Criteria
  - State-of-the-art success rates
  - Use Cases
  - Systems and tools
  - Available datasets and corpora
- 



# An explanation of the topic

- ▶ •Natural language processing (NLP) is a field of computer science to establish connection between computers and human languages
- ▶ •Named entity recognition is used for finding and classifying expressions in text into pre-defined categories, named entities (NE).
- ▶ •NE refers to real-world objects which are examples of person, location, organization, etc
- ▶ •Today, state-of-the-art NER systems for English scored up to 94% of F-Measure with recall and precision weighted equally while human experts scored about 97%



# The approaches and methods used to solve it

- ▶ Three approaches for Named Entity Recognition
  - ▶ Rule-base Named Entity Recognition
  - ▶ Machine Learning NER
  - ▶ Hybrid NER

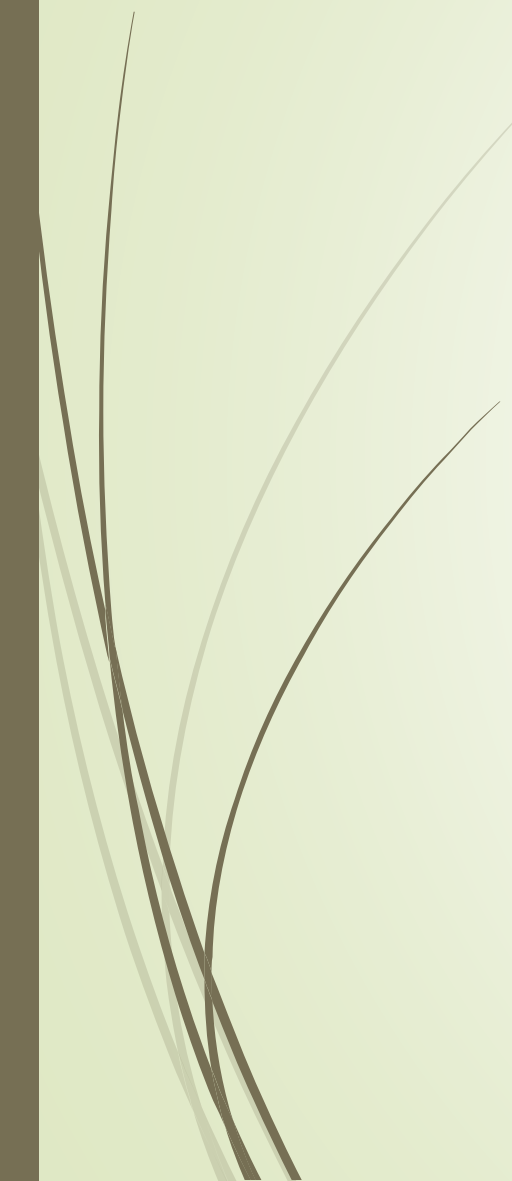
# Rule-Based Named Entity Recognition

- ▶ Predefined transformation rules;
  - ▶ Hand-crafted grammar rules
  - ▶ Gazetteers
  - ▶ Language dependent

$\langle \textit{Prof.}, \textit{Capitalized\_word}(X) \rangle \Rightarrow \textit{person\_named}(X)$

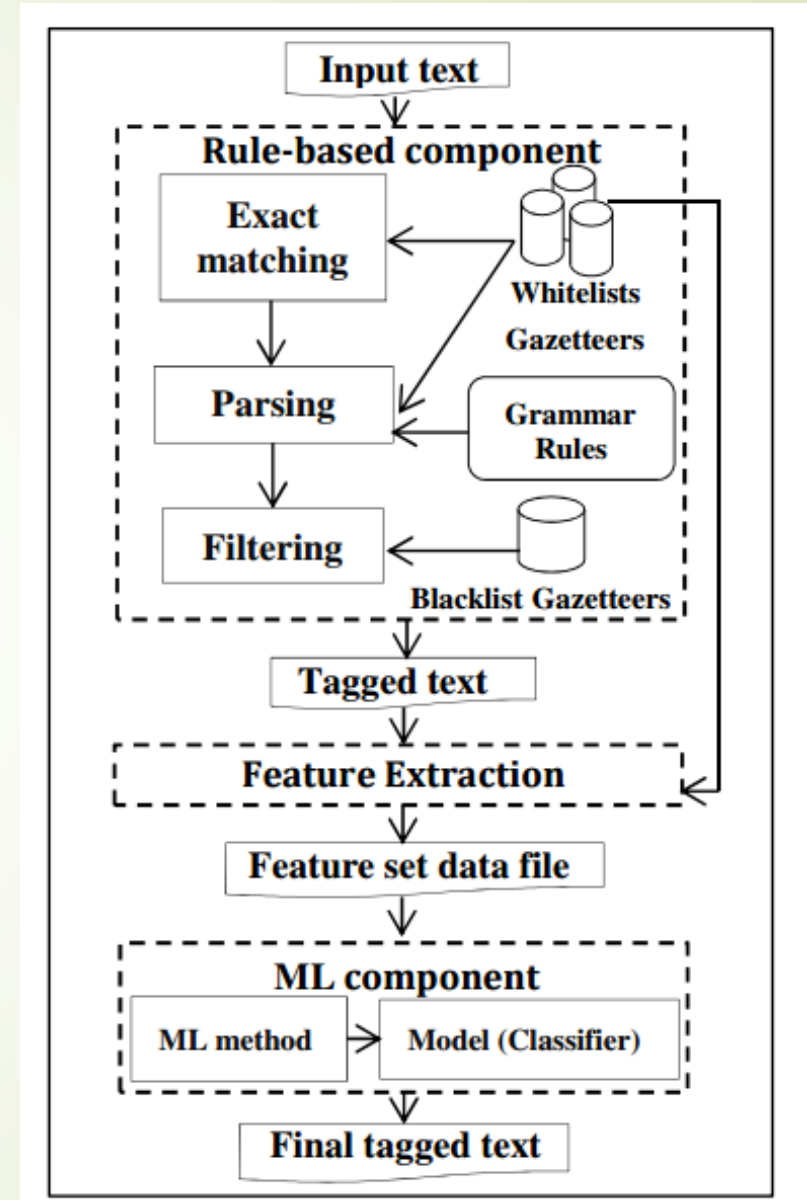


# Machine Learning Approach

- ▶ Entity recognition as a classification problem
  - ▶ Statistical models;
    - ▶ Conditional Random Fields
    - ▶ Maximum Entropy Markov Model
    - ▶ Support Vector Machine
    - ▶ Hidden Markov Models
- 

# Hybrid Approach

- Combination of rule-based and machine learning approaches
- 73 rules, 93 gazetteers 23.929 named entities
- 88.2% success rate





# How the approaches differ among different languages

## ► **Problems:**

- No standardization of written text - Arabic
- Ambiguity - Arabic
- Lots of variations exists in spelling writing style – Indian Languages
- Complex structure - Common
- Lack of resources - Common





# How the approaches differ among different languages

## ➤ **Used Methods:**

- Indian Language, Greek -> Rule based approach with adequate directory
- Arabic Language -> A hybrid system (Rule-based NER, Feature Engineering and ML-based NER)

# The success criteria used for evaluation

- Precision, Recall and F-Score to evaluate algorithms
- Recall is the fraction of relevant instances that are retrieved  $(TP/(TP + FN))$
- Precision is the fraction of retrieved instances which are relevant  $(TP / (TP + FP))$
- F-Score is the harmonic mean of precision and recall

		actual value		total
		<i>p</i>	<i>n</i>	
prediction outcome	<i>p'</i>	True Positive	False Positive	<i>P'</i>
	<i>n'</i>	False Negative	True Negative	<i>N'</i>
total		<i>P</i>	<i>N</i>	

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

# State-of-the-art success rates

- Hand-crafted Rule based approach

	<b>Organization</b>	<b>Person</b>	<b>Location</b>
<i>Precision</i>	0.898	0.875	0.905
<i>Recall</i>	0.842	0.765	0.756
<i>F-measure</i>	0.869	0.816	0.824

- A rule based approach by rule mining & Max Entropy

	Rule Association		Maximum Entropy	
	Recall	Precision	Recall	Precision
Dict	57.57	86.62	59.24	41.15
Bigram	34.37	93.21	57.40	65.03
Feature	44.84	67.75	49.56	58.99
Bigram+Dict	60.44	89.59	53.72	69.48
Feature+Dict	66.34	83.43	43.70	60.89
Bigram+Feature	53.73	77.61	59.61	76.10

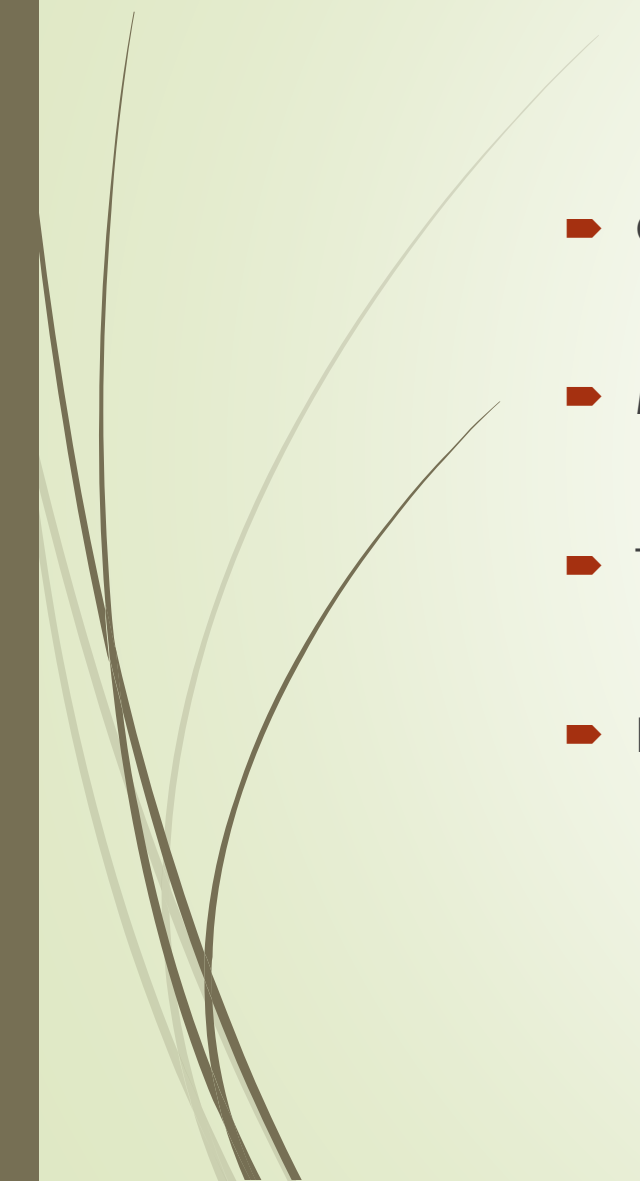
# State-of-the-art success rates

- ▶ A hybrid system

No.	Entity type	Precision (%)	Recall (%)	F-measure (%)
1	Person	86.3	89.2	87.7
2	Location	77.4	96.8	85.9
3	Company	81.45	84.95	83.15
4	Date	91.2	92.3	91.6
5	Time	97.25	94.5	95.4
6	Price	100	99.45	98.6
7	Measurement	97.8	97.3	97.2
8	Phone no.	94.9	87.9	91.3
9	ISBN	94.8	95.8	95.3
10	File name	95.7	97.1	96.4



# Example use cases

- ▶ Question Answering Systems
  - ▶ Machine Translation Systems
  - ▶ Text Mining
  - ▶ Bioinformatics
- 

# Systems and tools used currently

- ▶ Stanford Named Entity Recognizer (SNER)

Stanford CoreNLP

— Text to annotate —  
Nice weather in Turkey today.

— Annotations —  
named entities X

— Language —  
English

Submit

**Named Entity Recognition:**

1 Nice weather in **LOCATION** Turkey **DATE** 2017-05-02 today .

**CoreNLP Tools:**

TokensRegex Semgrep Tregex

Enter a **TokensRegex** expression to run against the above sentence:

e.g., (?foctype [{pos:JJ}]+ ) fox

Match

# Systems and tools used currently

## ➤ ITU NLP TOOL

### ITU Turkish Natural Language Processing Pipeline ITU Türkçe Doğal Dil İşleme Yazılım Zinciri

### Turkish Named Entity Recognizer

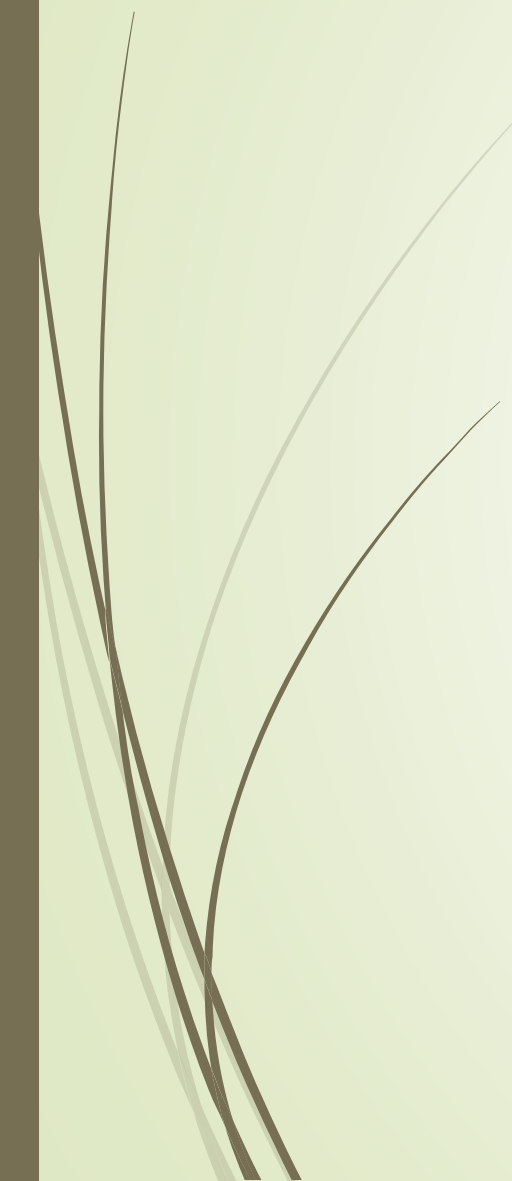
```
<DOC> <DOC>+BDTag  
<S> <S>+BSTag  
Bugün bugün+Noun+A3sg+Pnon+Nom bugün+Adverb  
İstanbul'da İstanbul+Noun+Prop+A3sg+Pnon+Loc  
hava hava+Noun+A3sg+Pnon+Nom hav+Noun+A3sg+Pnon+Dat  
güzel güzel+Adj güzel+Noun+NAdj+A3sg+Pnon+Nom güzel+Adverb  
</S> </S>+ESTag  
<DOC> <DOC>+EDTag
```

Send to Ner

```
<DOC> <DOC>+BDTag 0  
<S> <S>+BSTag 0  
Bugün bugün+Noun+A3sg+Pnon+Nom bugün+Adverb 0  
İstanbul'da İstanbul+Noun+Prop+A3sg+Pnon+Loc B-LOCATION  
hava hava+Noun+A3sg+Pnon+Nom hav+Noun+A3sg+Pnon+Dat 0  
güzel güzel+Adj güzel+Noun+NAdj+A3sg+Pnon+Nom güzel+Adverb  
</S> </S>+ESTag  
<DOC> <DOC>+EDTag 0
```



# Available data sets and corpora

- ▶ Training Data sets
  - ▶ Development Data sets
  - ▶ Test Data sets
- 



# Available data sets and corpora

- Reuters Ltd data collection
- RCV1 810.000 Reuters News stories in English

parent: None	child: Root	child-description: No Description
parent: CCAT	child: C11	child-description: STRATEGY/PLANS
parent: CCAT	child: C12	child-description: LEGAL/JUDICIAL
parent: CCAT	child: C13	child-description: REGULATION/POLICY
parent: CCAT	child: C14	child-description: SHARE LISTINGS
parent: CCAT	child: C15	child-description: PERFORMANCE
parent: C15	child: C151	child-description: ACCOUNTS/EARNINGS
parent: C151	child: C1511	child-description: ANNUAL RESULTS
parent: C15	child: C152	child-description: COMMENT/FORECASTS
parent: CCAT	child: C16	child-description: INSOLVENCY/LIQUIDITY
parent: CCAT	child: C17	child-description: FUNDING/CAPITAL
parent: C17	child: C171	child-description: SHARE CAPITAL
parent: C17	child: C172	child-description: BONDS/DEBT ISSUES
parent: C17	child: C173	child-description: LOANS/CREDITS
parent: C17	child: C174	child-description: CREDIT RATINGS
parent: CCAT	child: C18	child-description: OWNERSHIP CHANGES
parent: C18	child: C181	child-description: MERGERS/ACQUISITIONS
parent: C18	child: C182	child-description: ASSET TRANSFERS
parent: C18	child: C183	child-description: PRIVATISATIONS
parent: CCAT	child: C21	child-description: PRODUCTION/SERVICES
parent: CCAT	child: C22	child-description: NEW PRODUCTS/SERVICES
parent: CCAT	child: C23	child-description: RESEARCH/DEVELOPMENT
parent: CCAT	child: C24	child-description: CAPACITY/FACILITIES

# Available data sets and corpora

- ▶ CoNLL-2003
- ▶ Special Interest Group on Natural Language Learning (SIGNLL)
- ▶ Location, Person, Organization

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O



Thank You