Named Entity Recognition

Atil Vural & Dilek Kayahan
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
• 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

\[
\langle \text{Prof.}, \text{Capitalized\_word}(X) \rangle \Rightarrow \text{person\_named}(X)
\]
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 $\frac{TP}{TP + FN}$
- Precision is the fraction of retrieved instances which are relevant $\frac{TP}{TP + FP}$
- F-Score is the harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
State-of-the-art success rates

- Hand-crafted Rule based approach

- A rule based approach by rule mining & Max Entropy
State-of-the-art success rates

- A hybrid system

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

- Question Answering Systems
- Machine Translation Systems
- Text Mining
- Bioinformatics
Systems and tools used currently

- Stanford Named Entity Recognizer (SNER)
Systems and tools used currently

- ITU NLP TOOL

ITU Turkish Natural Language Processing Pipeline

Turkish Named Entity Recognizer
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
Available data sets and corpora

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