

Implementation of Name Entity Recognition System and its Evaluation

Group 28

Huang Yanzhen DC126732

Che Zirui DC127901

Introduction: To build a MaxEnt Model

Feature Selection

Select proper and precise features that can best help to distinguish a word entity from “others” and “name”.

1

Feature Weighting

Use the training methods to assign each feature an importance weight for further prediction.

2

Model Evaluation

The assigned values of weights determines the behavior of the model. The model predicts each word entity with the observed features and weights.

3



Feature Selection



**Frontend-Server
Model**



**Evaluations and
Demo**



Feature Selection

/ Feature Selection

8 custom features + 3 baseline features

Johnson
MacArthur
D.
...

Internal Pattern Features

Internal

Analyzes word patterns

Assumes that some specific pattern
could distinguish names

January
Tuesday
China
She

Library Features

Semi-Internal

Base on the experiences

Assumes that a word that is in some
class is likely or not likely to be
classified as a name

Attributive Clause
Start of Sentence
Positional Status

Contextual Features

External

Base on the contextual environment

Assumes that a word that has certain
contextual environment is likely or not
likely to be a name

/ Feature Selection / Internal Pattern Features

Johnson
MacArthur
D.
...

Internal Pattern Features

Internal

Analyzes word patterns

Assumes that some specific pattern could distinguish names

Feature Name		Description	Explanation	Match Examples
Positive Patterns	<i>p_cap_low</i>	Start with Capital Letter, and the rest of the letters are lowercase. There may be: - A prime after the first letter; - A second cap letter in the third or forth letter's space.	In English, names always start with a capital letter. There are some special styles that is quite unique in names.	Jonathan, Jason MacArthur, McDonald O'Brien
	<i>p_cap_period</i>	A single capital letter followed by a period.	In English, this pattern in most circumstances represent human name initials.	Donald J. Trump George W. Bush
Negative Patterns	<i>p_noun_like</i>	A word that has an ending like a noun. Specifically: -tion, -ment, -ness, -ship, -hood, -age, -ance, -ence	These suffixes are used to derive a noun from an adjective or adverb. These derivations are less likely to be names compared to other nouns.	movement, action, correctness, membership, likelihood, usage, allowance

/ Feature Selection / Library Features

January
Tuesday
China
She

Library Features

Semi-Internal

Base on the experiences

Assumes that a word that is in some class is likely or not likely to be classified as a name

Libraries		Python Package Name	Examples
Positive Library	Useful Names	<code>nltk.corpus.names</code>	James, Jonathan, ...
Negative Libraries	Week Names	<i>self-defined</i>	Tuesday, Wednesday, Thursday, ...
	Month Names	<i>self-defined</i>	January, February, March, April, ...
	Country Names	<code>geonamescache.countries</code>	China, Japan, United States, ...
	City Names	<code>geonamescache.cities</code>	London, Zhuhai, Hong Kong, Macau, ...
	Stopwords	<code>nltk.corpus.stopwords</code>	He, She, is, that, ...

/ Feature Selection / Contextual Features

Attributive Clause
Start of Sentence
Positional Status

Contextual Features

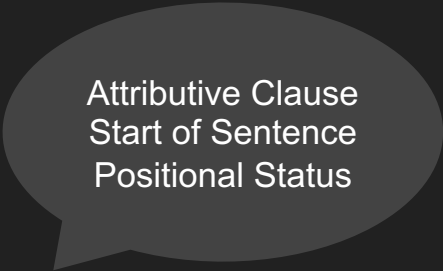
External

Base on the contextual environment

Assumes that a word that has certain contextual environment is likely or not likely to be a name

Feature Name		Description	Explanation
Positive Context	<i>is_start_of_sentence</i>	A word being at the start of a sentence. - This word has position of 0. - This word is after a concrete period.	It is highly likely that a word entity that fits the pattern of a name defined before is the start of the sentence.
	<i>is_target_of_clause</i>	Is the target of the restricted attributive clause.	We often refer someone with addition informations using restricted attributive clause. For instance: <i>Jane, who was my friend, went to the park.</i> This clause puts high probability to the target entity that it is a name.
	<i>is_after_status</i>	Is after the social status in English, like Mr., Ms.	It is very common to put names after social statuses.

/ Feature Selection / Contextual Features / Clause



Attributive Clause
Start of Sentence
Positional Status

Attributive Clause Examples

Donald J. Trump, **who** was a former US president, was a successful business man.

Michael Rosen, **whose** son died before him, wrote *We're Going on a Bear Hunt*.

Contextual Features

External

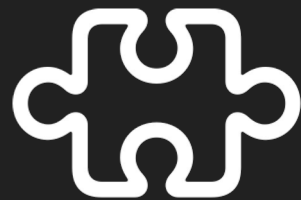
Base on the contextual environment

Assumes that a word that has certain contextual environment is likely or not likely to be a name

Clause Form

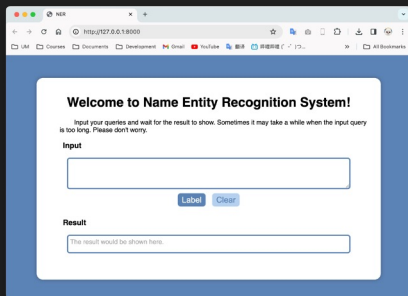
<*Entity*>, who/whose <*verb-phrase*>.

Highly Likely to be a name!



**Frontend-Server
Model**

/ Frontend-Server Model

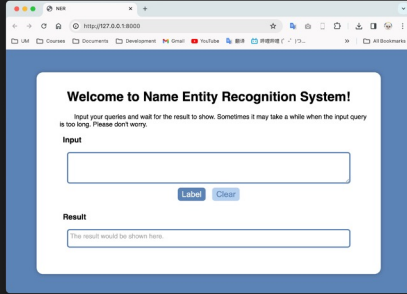


Frontend Framework
HTML + CSS + JS



Backend Framework
Django + Python

/ Frontend-Server Model / Frontend Framework



Frontend Framework
HTML + CSS + JS

HTML + CSS:

- Structure & Style of website.

JavaScript:

- Defines how the input query is submitted to the backend (XMLHttpRequest, etc).
- Define how the response from the server is handled.

/ Frontend-Server Model / Backend Framework



Backend Framework

Django + Python

Django: Model-Template-View Framework

View: Defines how the backend respond to the frontend, facing some specific requests.

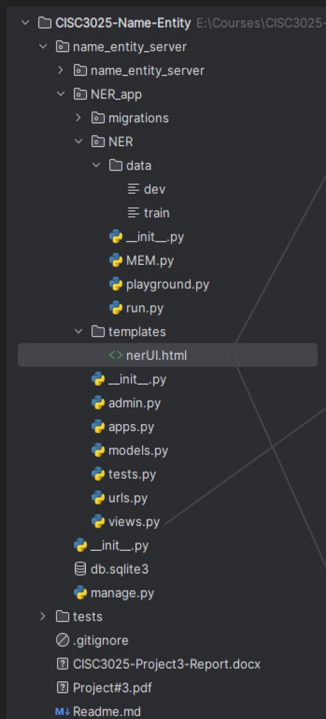
In this case:

- The predicted query of the MaxEnt Model is the response from the backend.
- The backend responds the frontend in JSON format.

/ Frontend-Server Model / Backend Framework



Backend Framework
Django + Python



1

```
// Send form data.  
document.getElementById('result').classList.add('empty-field');  
document.getElementById('result').innerHTML = "Processing...";  
xhr.send(new URLSearchParams(formData).toString());
```

2

```
1 usage  ± Yanzen Huang +1  
def resultView(request):  
    input_query = request.POST.get("input-query", "<blank>")  
  
    # Add your name entity processing here!!!  
    # Please never alter the path!!!!  
    model_pkl_path = os.path.abspath('name_entity_server/static/model.pkl').replace(_old: '\\', _new: '/')  
  
    # Get modified name string  
    names, labels = playground.predict(input_query, MEM, model_pkl_path)  
    output_query = (  
        names  
        # + " <br> "  
        # + labels  
    )  
  
    # output_query = input_query + " ---- from backend"  
    return JsonResponse({"result": output_query})
```

Invoke Model

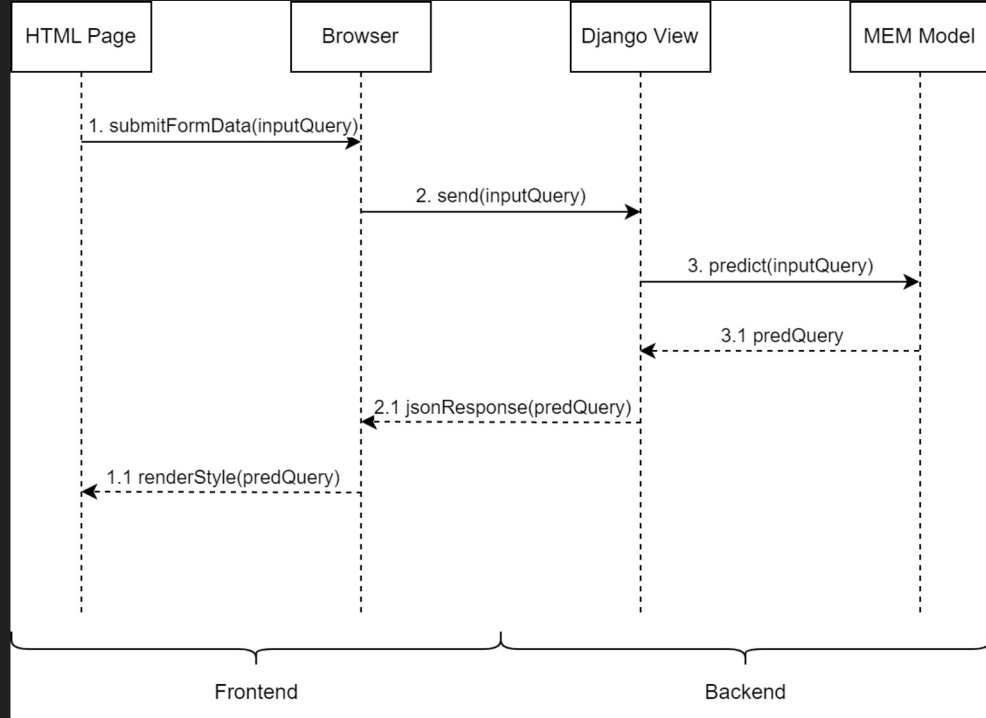
3

```
// Async. Register a response Handler event listener.  
xhr.onload = function() {  
    if (this.status === 200) {  
        let response = JSON.parse(this.responseText);  
        document.getElementById('result').innerHTML = response.result;  
        document.getElementById('result').classList.remove('empty-field');  
    }  
};
```

/ Frontend-Server Model / Backend Framework



Backend Framework
Django + Python





Evaluations and Demo

/ Evaluations and Demo /

Final Training Result

```
Training classifier...
==> Training (5 iterations)

  Iteration   Log Likelihood   Accuracy
-----
      1         -0.69315         0.055
      2         -0.09506         0.946
      3         -0.07840         0.967
      4         -0.06727         0.976
    Final         -0.05937         0.981

<ConditionalExponentialClassifier: 2 labels, 23877 features>
PS E:\Courses\CISC3025-Name-Entity\name_entity_server\NER_app\NER>
```

Final Testing Result

```
Testing classifier...
f_score=      0.9202
accuracy=     0.9752
recall=       0.8145
precision=    0.9609

PS E:\Courses\CISC3025-Name-Entity\name_entity_server\NER_app\NER>
```

Final Show Examples

```
PS E:\Courses\CISC3025-Name-Entity\name_e
Words          P(PERSON)  P(0)
-----
EU              0.0126   *0.9874
rejects         0.0426   *0.9574
German          0.0202   *0.9798
call            0.0426   *0.9574
to              0.0087   *0.9913
boycott         0.0426   *0.9574
British         0.0203   *0.9797
Lamb            0.0426   *0.9574
.               0.0087   *0.9913
Peter           *0.5999    0.4001
Blackburn       *0.3352    0.6648
BRUSSELS        0.1707   *0.8293
1996-08-22      0.0431   *0.9569
The             0.0206   *0.9794
European        0.1043   *0.8957
Commission      0.1043   *0.8957
said            0.0385   *0.9615
on              0.0087   *0.9913
Thursday        0.0202   *0.9798
it              0.0087   *0.9913
disagreed       0.0426   *0.9574
with            0.0087   *0.9913
```

What we think

- Unbelievably “Good”
- Recall too high, may indicate more overfitting

/ Evaluations and Demo / Eliminate Features

Feature Name		Description
Negative Internal Pattern Features	<i>p_name_prefix</i>	Social Status
	<i>p_possessive_like</i>	Possessive case of a pronoun.
	<i>p_country_abbrev_like</i>	Abbreviation of country names. Like U.K. or U.S.
	<i>p_num_slash</i>	A set of numeric descriptions. For instance, 12-20
	<i>is_posessive</i>	Is before the entity "'s".

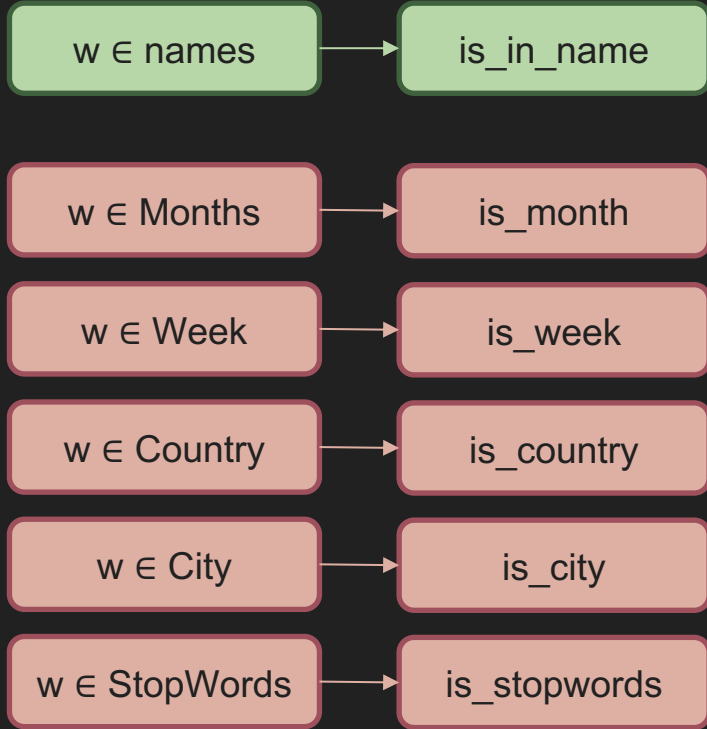
Feature Name		Description
Negative Contextual Features	<i>is_around_first</i>	Social Status
	<i>is_last_word</i>	Last word in sentence
	<i>is_after_name_prefix</i>	After social statuses, like Mr., Ms.
	<i>is_posessive</i>	Is before the entity "'s".
	<i>is_after_verb</i>	A verb is after it.

/ Evaluations and Demo / Eliminate Features

Too much Negative Internal Pattern Features.

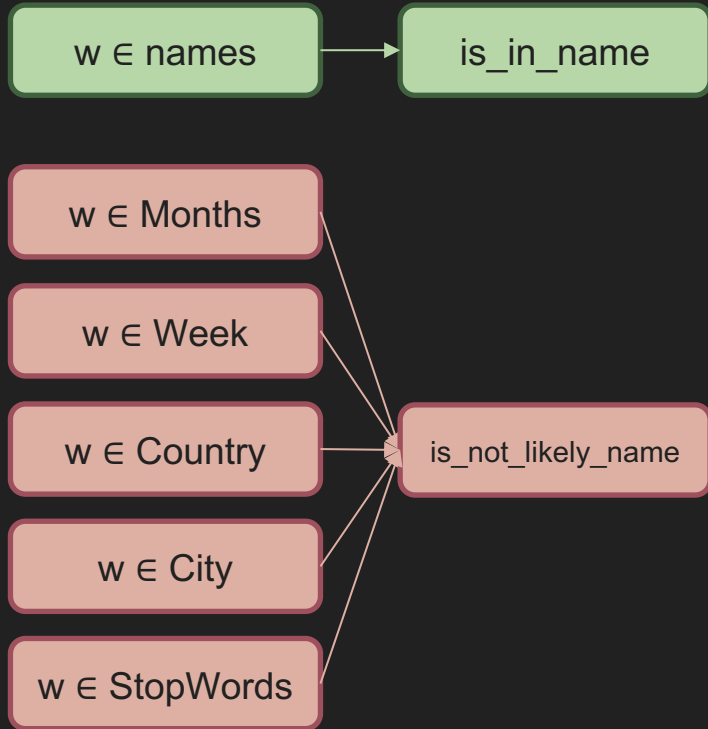
- Helps to improve F-Score, especially recall
- However, causes over-fitting.
- Discourage innovation.
- Synonym: You can't make a child successful by regulating him too much.

/ Evaluations and Demo / Merge Library Features



- Library Features were dispersed at first
- That is, one library matches to one feature
 - However, again, we don't want to tell the model too much about "what not to do", instead of what to do.
 - Therefore, these features are merged into one.

/ Evaluations and Demo / Merge Library Features



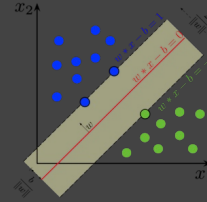
- Library Features were dispersed at first
- That is, one library matches to one feature
 - However, again, we don't want to tell the model too much about "what not to do", instead of what to do.
 - Therefore, these features are merged into one.

/ Evaluations and Demo / Future Improvements

POS

POS Features

Part of Speech Features are quite effective for discriminating entity classes .



Pre-Processing

Pre-process training data to eliminate noise, using Support Vector Machine.

/ Evaluations and Demo /

Live Demo

Thanks for Listening!