Week 4: Named Entity Recognition (Cont.)

Nhung Nguyen
slides courtesy of NaCTeM

Machine learning-based approaches

NER as a tagging problem (BIO scheme)

	Adam	Smith	works	for	IBM	in	London	.
POS tagging	NNP	NNP	VBZ	IN	NNP	IN	NNP	.
Entity identification	B	I	O	O	B	O	B	O
Entity recognition	B_PER	I_PER	O	O	B_ORG	O	B_LOC	O

Begin the mention

Inside the mention

- # classes = 2 * # entity types + 1

Classification approach

- Predicting tags

  `probability(tag|token) = function f`

- Local approach: tags are ***independent*** each other
  - Any classifiers for sequence can be used, e.g., RNN, LSTM, BiLSTM

- Global approach: tags are ***dependent*** each other
  - Hidden Markov Model (HMM)
  - Conditional Random Fields (CRF)

Conditional Random Fields

Sequence model

- Relax the independence assumption by arranging the output variables in a linear chain
- Hidden Markov Model (HMM):
  - A sequence of input: $X = \{x_t\}_{t=1}^{T}$
  - A sequence of states: $Y = \{Y_t\}_{t=1}^{T}$
  - y_t is only dependent on the previous state y_{t-1}
  - x_t is only dependent on the current state y_t
- Joint distribution:

$$p(y, x) = \Pi_{t-1}^{T} p(y_{t-1} | y_t) p(x_t | y_t)$$

Conditional Random Fields (CRFs)

- a discriminative sequence model for sequence labelling
- finds the *most probable label sequence* $\boldsymbol{y}^*$ given an observation sequence $\boldsymbol{x}$

$$y^* = \text{argmax}_y p(y|x)$$

where $\boldsymbol{x}$ consists of the sequence of tokens from input text

Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282–289.

Linear-chain CRFs

Computation of probability

weight   feature function

$$p(y|x) = \frac{1}{Z_x}\exp(\Sigma_{t=1}^{T}\Sigma_{k=1}^{K} w_k f_k(y_t, y_{t-1}, x_t))$$

summation over all tokens

summation over all feature functions

Normalisation factor: to make sure the sum of probability is equal to 1

$$Z_x = \Sigma_y \exp(\Sigma_{t=1}^{T}\Sigma_{k=1}^{K} w_k f_k(y_t, y_{t-1}, x_t))$$

Feature function

- Characterises the input

$$f(y_t, y_{t-1}, x_t) = \begin{cases} 1, \text{if 1st letter of } x_t \text{is uppercase} \\ 0, \text{otherwise} \end{cases}$$

- Example

y_{t-1}= O, y_t= B-PERSON, 1st letter of x_t is uppercase

Feature types

- Contextual
  - current word *wo*
  - words around *wo* in [-3,…,+3] window
- Part-of-speech tag (when available)
- Trigger words
  - for person (Mr, Miss, Dr, PhD)
  - for location (city, street)
  - for organisation (Ltd., Co.)

Feature types (Cont.)

- Length (in terms of number of tokens)
- Orthographic (binary and not mutually exclusive)
  - initial-caps, all-caps, lonely-initial
  - all-digits contains-dots, punctuation-mark
  - single-char, contains-hyphen, URL
  - roman-numeral
- Suffixes (length 1 to 4)
  - each component of the NE
  - whole NE

Feature types (Cont.)

- Gazetteers
  - geographical locations
  - first names, surnames
  - company names
  - many others
  - whole NE is in gazetteer?
  - any component of the NE appears in gazetteer?

The more useful features you incorporate, the more powerful your learner gets!

Examples of features: Contextual

- current word w_o
- words around w_o in [-3,…,+3] window

w_0	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam						
Smith						
works						
for						
IBM						
in						
London						
.						

Examples of features: Contextual

- current word w_o
- words around w_o in [-3,…,+3] window

w_0	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam	null					
Smith	null					
works	null					
for	Adam					
IBM	Smith					
in	works					
London	for					
.	IBM					

Examples of features: Contextual

- current word w_o
- words around w_o in [-3,…,+3] window

w_0	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam	null	null				
Smith	null	null				
works	null	Adam				
for	Adam	Smith				
IBM	Smith	works				
in	works	for				
London	for	IBM				
.	IBM	in				

Examples of features: Contextual

- current word w_o
- words around w_o in [-3,…,+3] window

w_0	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam	null	null	null			
Smith	null	null	Adam			
works	null	Adam	Smith			
for	Adam	Smith	works			
IBM	Smith	works	for			
in	works	for	IBM			
London	for	IBM	in			
.	IBM	in	London			

Examples of features: Contextual

- current word w_o
- words around w_o in [-3,…,+3] window

w_0	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam	null	null	null	Smith	works	for
Smith	null	null	Adam	works	for	IBM
works	null	Adam	Smith	for	IBM	in
for	Adam	Smith	works	IBM	in	London
IBM	Smith	works	for	in	London	.
in	works	for	IBM	London	.	null
London	for	IBM	in	.	null	null
.	IBM	in	London	null	null	null

Examples of features: Orthographic

- Is initial letter capitalised?
- Are all letters capitalised?

w_0	InitC	AllC	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam			null	null	null	Smith	works	for
Smith			null	null	Adam	works	for	IBM
works			null	Adam	Smith	for	IBM	in
for			Adam	Smith	works	IBM	in	London
IBM			Smith	works	for	in	London	.
in			works	for	IBM	London	.	null
London			for	IBM	in	.	null	null
.			IBM	in	London	null	null	null

Examples of features: Orthographic

- Is initial letter capitalised?
- Are all letters capitalised?

w_0	InitC	AllC	w_{-3}	w_{-2}	w_{-1}	w_1	w_2	w_3
Adam	1	0	null	null	null	Smith	works	for
Smith	1	0	null	null	Adam	works	for	IBM
works	0	0	null	Adam	Smith	for	IBM	in
for	0	0	Adam	Smith	works	IBM	in	London
IBM	1	1	Smith	works	for	in	London	.
in	0	0	works	for	IBM	London	.	null
London	1	0	for	IBM	in	.	null	null
.	0	0	IBM	in	London	null	null	null

Pros vs. Cons

- Pros:
  - Features are intuitive
  - It is easy to interpret and debug the model
  - High performance

The solution: Neural Network!

- Cons:
  - Feature engineering → domain knowledge