Sequence Tagging

Computational Linguistics

Emory University

Jinho D. Choi

Part-of-Speech Tagging

Classify the part-of-speech tag of each token.

Jinho	is	a	professor
noun	verb	det.	noun
proper	3rd, present		common

https://github.com/emory-courses/cs571/blob/master/course/sequence_tagging.md#penn-pos-tagset

Supervised NLP

1. Collect
2. Train
3. Evaluate

a. Design a processing algorithm.

b. Extract (label, features) pairs.

c. Vectorize labels and features.

d. Build statistical models.

Feature Extraction

Extract the label and the features given the current state.

$$\{w_i, w_{i-1}, w_{i+1}, p_{i-1}\}$$

John/NNP is/VBZ a/DT teacher/NN

Label	F_0	F_1	F_2	F_3
NNP	John	∅	is	∅
VBZ	is	John	a	NNP
DT	a	is	teacher	VBZ
NN	teacher	a	∅	DT

John/NNP was/VBD a/DT student/NN

NNP	John	∅	was	∅
VBD	was	John	a	NNP
DT	a	was	student	VBD
NN	student	a	∅	DT

Feature Extraction

Filter out ones whose frequencies ≤ cutoff.

cutoff = 1

Label	F_0	F_1	F_2	F_3
NNP	John	∅	is	∅
VBZ	is	John	a	NNP
DT	a	is	teacher	VBZ
NN	teacher	a	∅	DT
NNP	John	∅	was	∅
VBD	was	John	a	NNP
DT	a	was	student	VBD
NN	student	a	∅	DT

Count

Label	{NNP:2, VBZ:1, DT:2, NN:2, VBD:1}
F_0	{John:2, is:1, a:2, teacher:1, was:1, student:1}
F_1	{John:2, is:1, a:2, was:1}
F_2	{is:1, a:2, teacher:1, was:1, student:1}
F_3	{NNP:2, VBZ:1, DT:2, VBD:1}

Feature Extraction

Assign an unique ID to each label and feature.

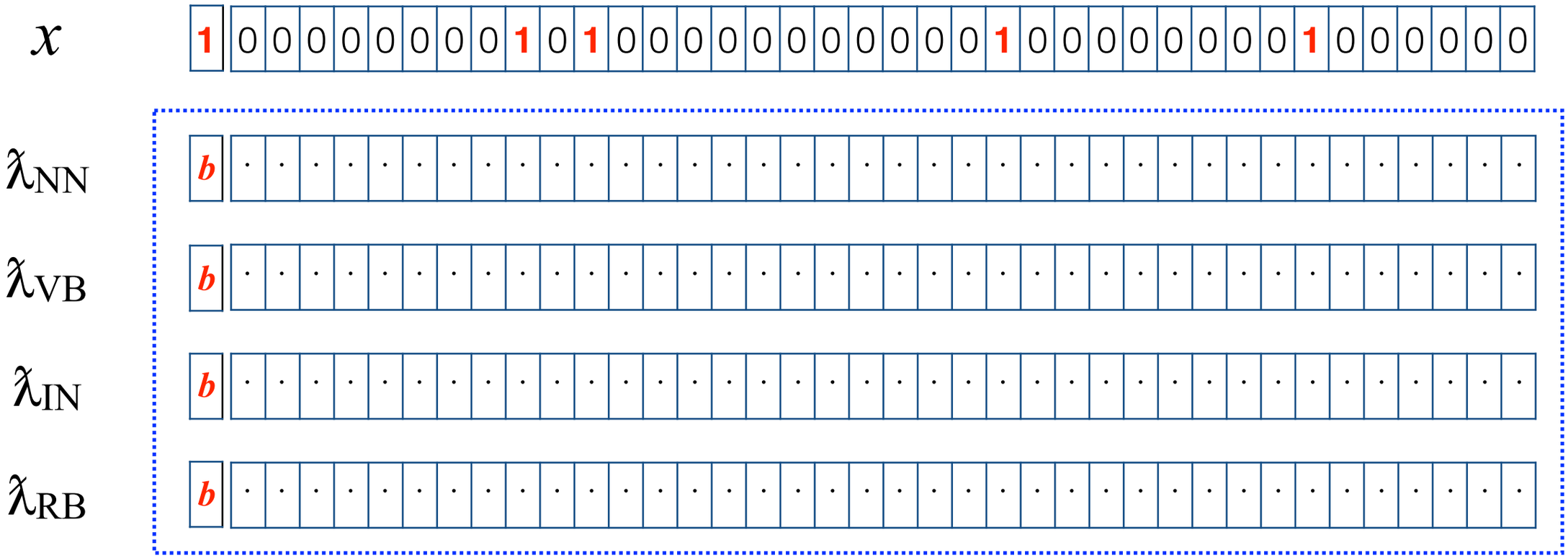
Label	**F_0**	**F_1**	**F_2**	**F_3**
NNP	John	∅	is	∅
VBZ	is	John	a	NNP
DT	a	is	teacher	VBZ
NN	teacher	a	∅	DT
NNP	John	∅	was	∅
VBD	was	John	a	NNP
DT	a	was	student	VBD
NN	student	a	∅	DT

Label	{NNP:0, DT:1, NN:2}
F_0	{John:1, a:2}
F_1	{John:3, a:4}
F_2	{a:5}
F_3	{NNP:6, DT:7}

0	1	1	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0
2	1	0	0	0	1	0	0	1
0	1	1	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0
2	1	0	0	0	1	0	0	1
	0	1	2	3	4	5	6	7

Softmax Regression

$$p(y|X) = \frac{1}{Z(\mathrm{x})} \exp\left\{ \overset{?}{\boxed{\lambda_y}} + \sum_{\forall k} \boxed{\lambda_{y,k}} \cdot \boxed{x_k} \right\}$$



$$p(y|X) = \frac{1}{Z(\mathrm{x})} \exp\left\{ \sum_{\forall k} \lambda_{y,k} \cdot x_k \right\}$$

Ambiguity Classes

The likely part-of-speech tag.

John ← NNP:100 NNP

study ← VB: 50, NN:50 VB_NN

interest ← JJ:70, NN:30 JJ or JJ_NN

Collect the ambiguity classes before training.

Use them as extra features.

Named Entity Recognition

Classify the named entity tag of each chunk.

Peson			Organization				Location		
Jinho	is	a	professor	at	Emory	University	in	the	United	States	of	America	.
PER	O	O	O	O	ORG	ORG	O	O	LOC	LOC	LOC	LOC	O

A chunk can be decomposed into a sequence of tokens.

Classify the named entity tag of each token.

Different from part-of-speech tagging?

BIO Notation

Peson Organization Location

Jinho is a professor at Emory University in the United States of America.

PER O O O O ORG ORG O O LOC LOC LOC LOC

Semantic overload Semantic overload

B-PER **B**-ORG **I**-ORG **B**-LOC **I**-LOC ...

B: Beginning
I: Inside
O: Outside

Still not enough?

BILOU Notation

B: Beginning
I: Inside
O: Outside
L: Last
U: Unit

Jinho	**B**-PER	**U**-PER
Emory	**B**-ORG	**B**-ORG
University	**I**-ORG	**L**-ORG
United	**B**-LOC	**B**-LOC
States	**I**-LOC	**I**-LOC
of	**I**-LOC	**I**-LOC
America	**I**-LOC	**L**-LOC

Features

Similar to part-of-speech tagging.

Features from knowledge-base.

Freebase: http://www.freebase.com

DBPedia: http://wiki.dbpedia.org

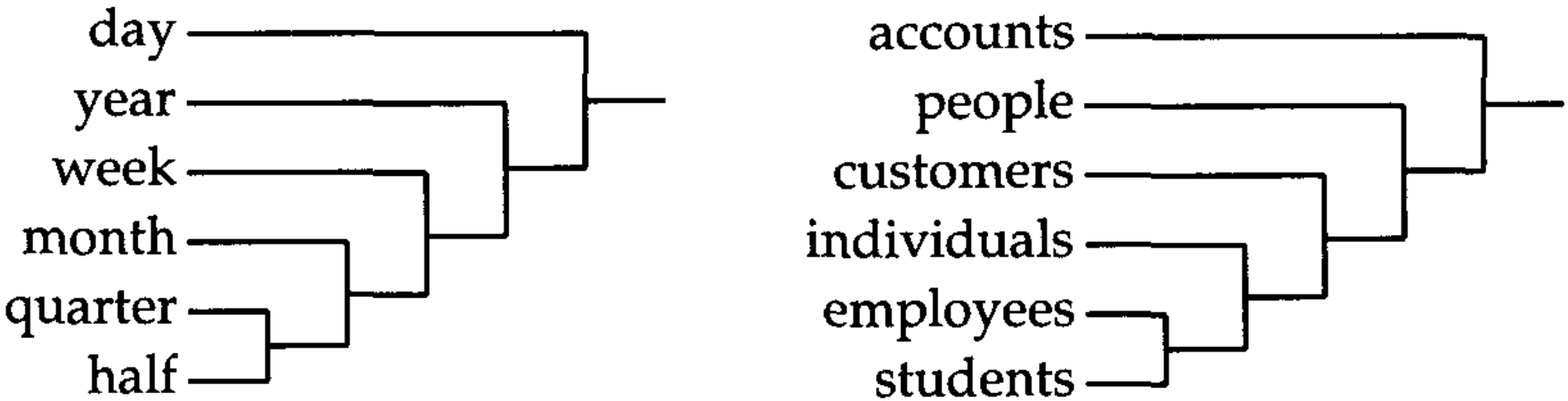
DBPedia Spotlight

https://dbpedia-spotlight.github.io/demo/

Clustering Features

Given a large corpus, construct word clusters.

Brown cluster



Use the cluster info as an extra feature for each token.

Evaluation

Jinho is a professor at Emory University in the United States of America.

U-P O O O O B-R L-R O O B-L I-L I-L L-L

↑Gold System↓

U-P O O O O U-R U-R O O B-L I-L L-L O

Exact match

Precision

$$p = \frac{\text{correct entities}}{\text{predicted entities}} = \frac{1}{4}$$

Recall

$$r = \frac{\text{correct entities}}{\text{true entities}} = \frac{1}{3}$$

$$F1 = 2 \cdot \frac{p \cdot r}{p + r}$$