

Information Extraction

ดร.อรรถพล ชั่รังษ์ตันติพิธี

August 14, 2019

Data ข้อมูล → Information ความรู้



lingling (๑) リン

16 127 559

★★★★★ Quality Review

ร้านในตำนานอักษรศาสตร์ จุฬาฯ

ขอยกให้ร้านนี้เป็นร้านในตำนานอักษรศาสตร์ จุฬาฯ ^^ เราทานมาตั้งแต่สมัยเรียนปริญญาตรี เมื่อตอนนานมาก มาแล้ว และห่างหายไปนาน จนล่าสุดมาเจอกันอีกครั้ง กับ NOW Food Delivery โดยบังเอิญ เลยลองสั่งมาทานค่ะ

เซ็ตที่สั่งมาเป็นข้าวเหนียวไก่ทอด+ข้าวเหนียว ราคาตามในแอปคือ 33 บาท ใส่ในกล่องกระดาษวัสดุอย่างดี แพ็กมาสวยงามน่าทานตามรูปเลยค่ะ โดยรสชาติความอร่อยยังเหมือนเดิม ทำให้หายคิดถึงไปได้มากเลยค่ะ

เมนูที่แนะนำโดยสมาชิก



ข้าวเหนียวไก่ทอด 9

ข้าวเหนียวไก่ทอด+เอ็นไก่ 2

ส้มตำไข่เค็ม 1

ยำไก่ทอด 1

ข้าวเหนียวเนื้อ 1

ดูกันหมด »

ข้าวเหนียวหมู-ไก่ 1

ภาษาจัดเป็นข้อมูลแบบไม่มีโครงสร้าง

- เปิดเพลงอะไรก็ได้ของปัล์มมิตอนหกโมงเย็น

Artist ID	Artist Name
1	Atom ชนกันต์
2	<u>Palmy</u>
3	Stamp
4	แจนจัง



Unstructured Data



Rockbox Brick เป็นลำโพงไร้สายที่มีเบสขนาดใหญ่ที่มีรูปร่างอิฐแบบคลาสสิก

ใช้บลูทูธเพื่อเชื่อมต่อแบบไร้สายกับอุปกรณ์ สามารถเชื่อมต่อเข้ากัน โทรศัพท์ แท็บเล็ตหรือโน้ตบุ๊ค สามารถเป็น Powerbank แบตเตอรี่ที่มีกำลังไฟ 4000 mAh ฟังเพลงต่อเนื่อง 20 ชั่วโมงจากการชาร์จไฟครึ่งเดียวไฟ

ขนาด 15.5 x 5.9 x 5.9 ซม.

ในชุดประกอบด้วย

Rockbox Brick

Micro-USB charging cable

3.5 mm audio cable พบสินค้าเพิ่มเติมจาก FRESHN REBEL

- สี : Indigo
- กำลังไฟฟ้า (วัตต์) : 4000 mAh

Structured Data

Brand = FRESHN REBEL

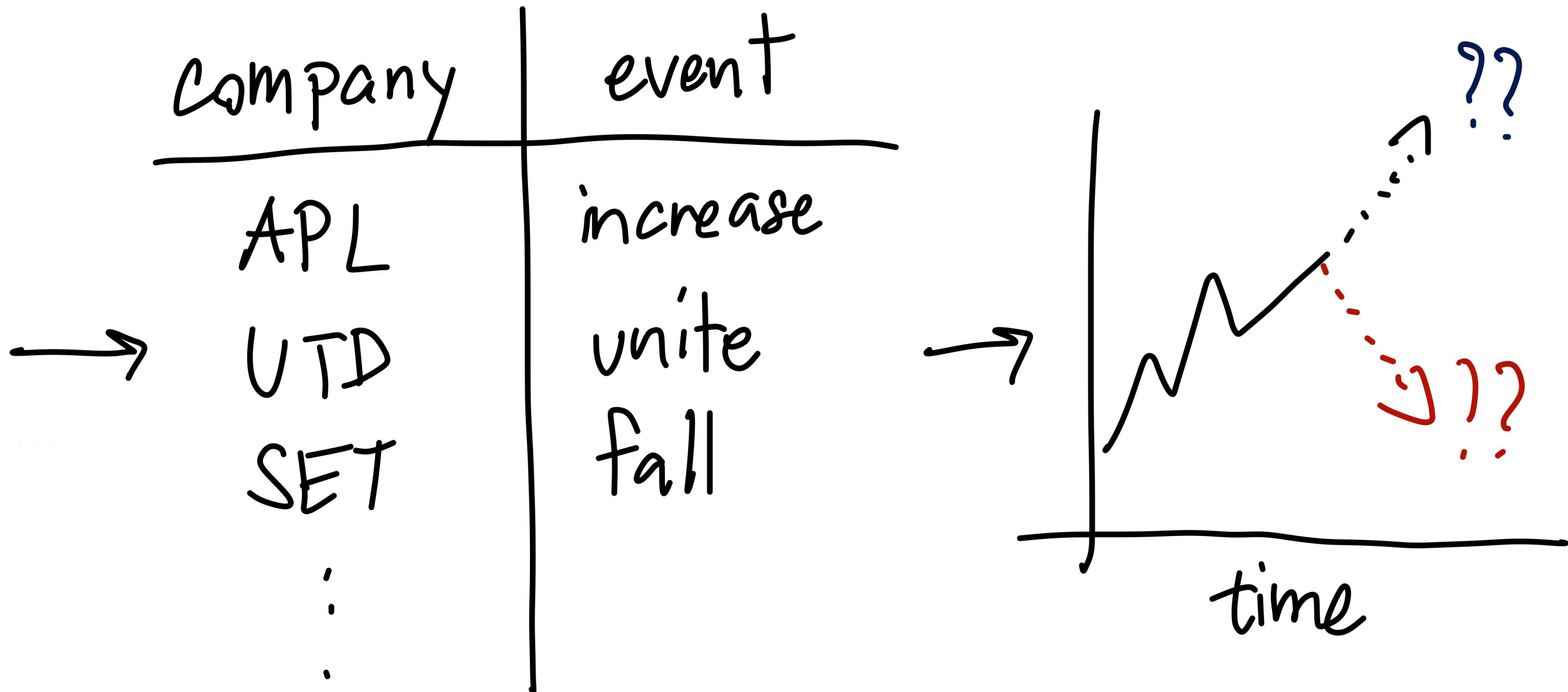
Color = Indigo

Type = Portable Speaker

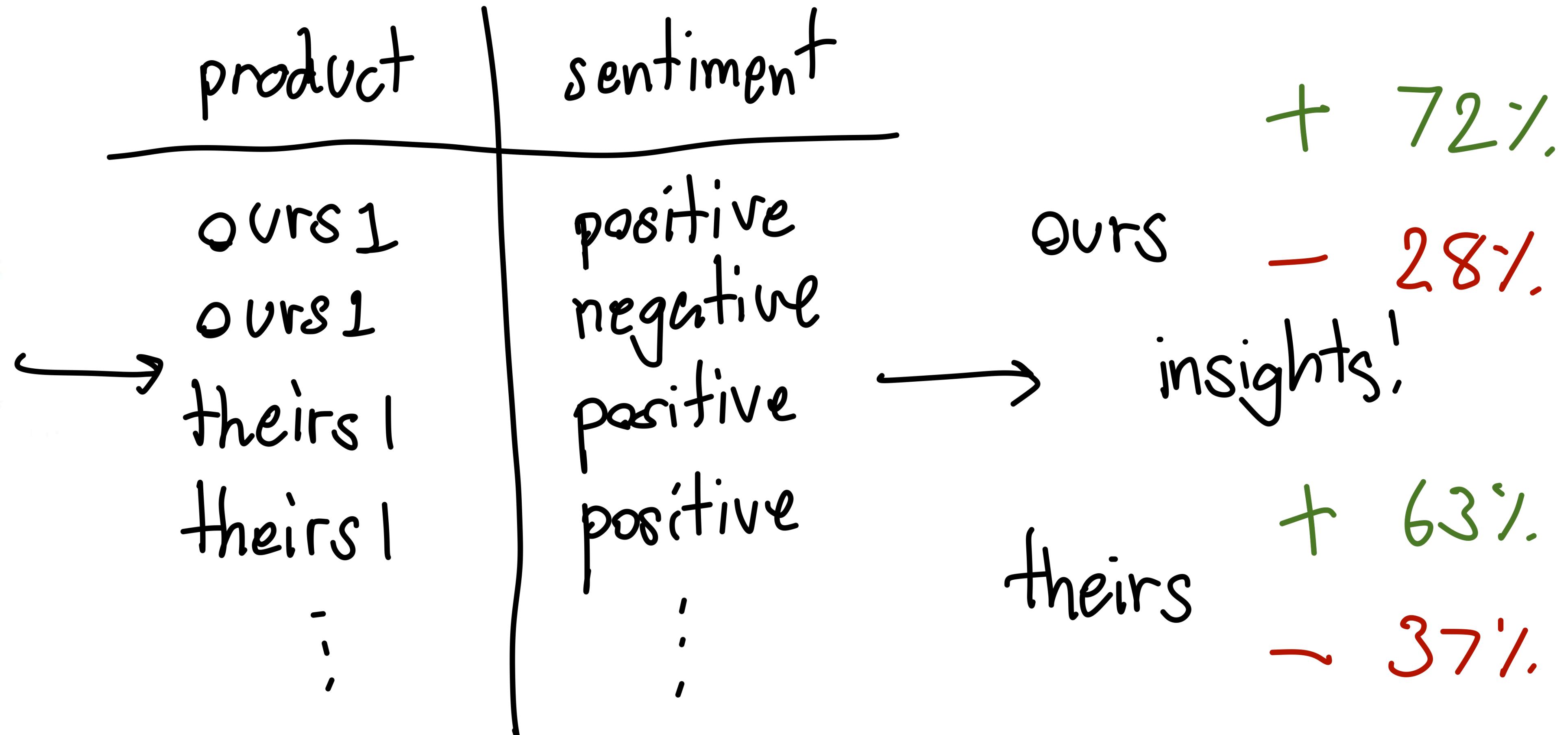
Bluetooth Speaker

Home Decor

Finances



Marketing + Brand Monitoring



Election Forecast



Where candidate sentiment		
		neutral
BKK	8	neutral
BKK	8	positive
BKK	6	positive
CNX	7	neutral
CNX	1	negative
:	:	:

→ campaign at CNX

Drug Administration



patient	time	event	diseases
1	0	x-ray	-
1	1	drug1	pneumonia
2	0	MR1	-
2	1	x-ray	-
2	2	drug2	infection

ສັດວະໄໄດ້ບ້າງ

- ຂໍ້ອຄນ ສພານທີ່ທາງກູມມີສາສຕ່ຣ໌ ຮ້ານຄ້າ ຂໍອອງຄໍກຣ
- ຂໍ້ອເພລັງ ຂໍ້ອສືລປິນ ຂໍ້ອວຳລົ້ມ
- ວັນ ເວລາ ວັນທີ ແທງກາຣນ໌
- ຍິນສ් ໂປຣຕິນ ຂໍ້ອຍາ ອາກາຣທາງແພທຍ໌ ເຄຣືອງມືອກາຣວິນິຈຊຍ໌ ຂໍ້ອເຫຼື້ອໂຮຄ
ຂໍ້ອໂຮຄ

การสกัดความรู้ (Information Extraction)

- การเปลี่ยน unstructured data (data ที่เป็น text นำไปใช้ได้ยาก)
เป็น structured data (data ที่เป็นตารางสามารถนำไปใช้ง่าย)

Part-of-Speech Tagging

ดีแทค ลด ค่าโหวร

ดีแทค

ลด

ค่าไฟร

สำหรับ

ลูกค้า

ใหม่

Universal POS Tag

Open-class words

- ADJ
- ADV
- INTJ
- NOUN
- PROPN
- VERB

Closed-class words

- ADP
- AUX
- CCONJ
- DET
- NUM
- PART
- PRON
- SCONJ

ทรัพย์ ดึง ลูกค้า ที่ ไม่ ชอบ ดีแทค

ทรู ดิจ ลูกค้า ที่ ไม่ ชอบ ดีแทค

Part-of-Speech Tagging + Base NP

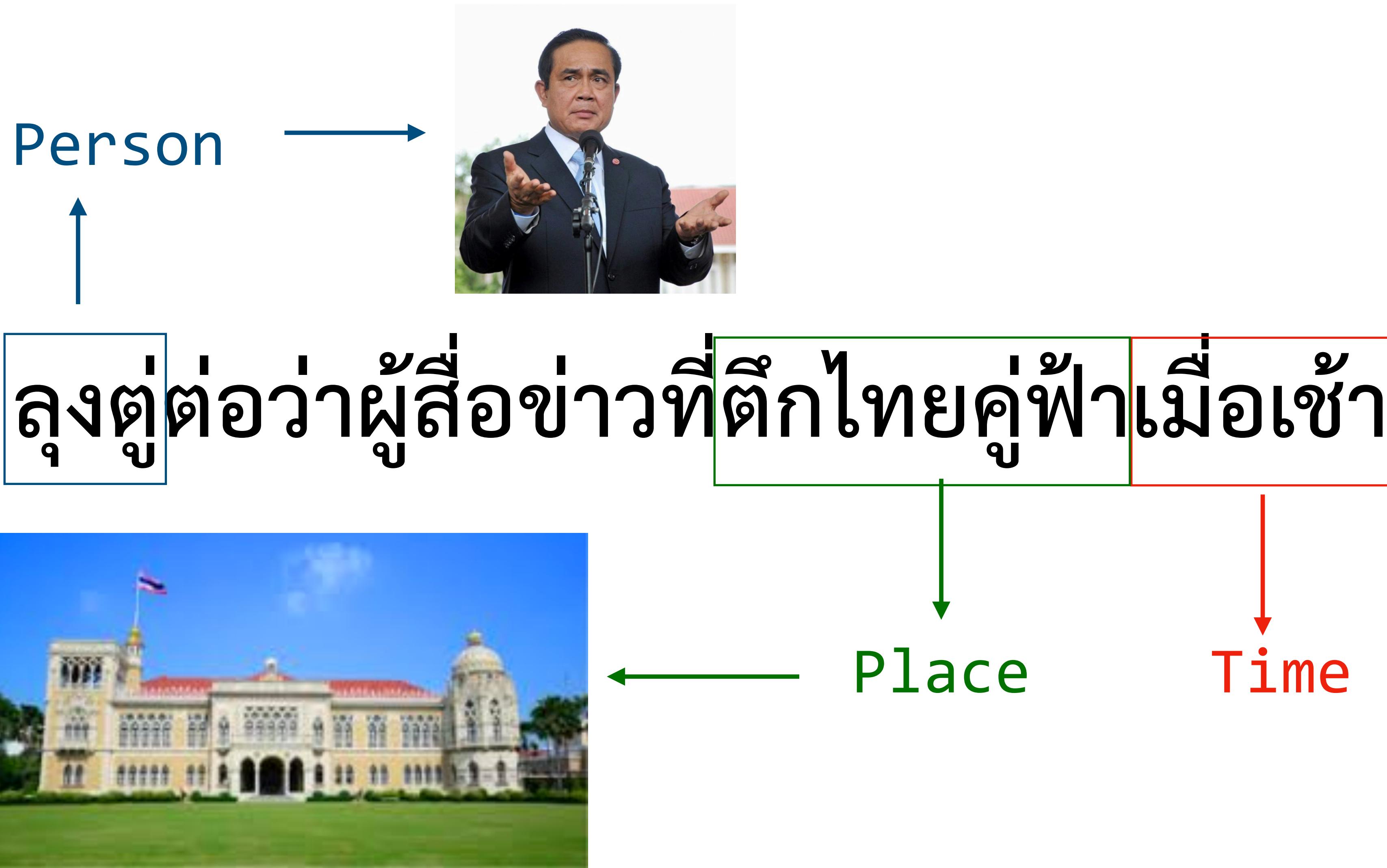
- Sequence labeling task
- การแบ่งนิດของคำทำให้เราเข้าถึงความหมายได้ระดับนึง
- Base NP Chunking ช่วยสกัดความรู้เกี่ยวกับ คน สัตว์ สิ่งของ
สถานที่ และสิ่งนามธรรมอื่นๆ

Named-Entity Recognition (NER)

displaCy Named Entity Visualizer

When Sebastian Thrun **PERSON** started working on self-driving cars at Google **ORG** in 2007 **DATE**, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American **NORP** car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun **PERSON**, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode **ORG** earlier this week **DATE**.

Named-Entity Recognition



NER Task Formulation

ลุงตู่ต่อว่าผู้สืบขอข่าวที่ตึกไทยคุ่ฟ้าเมื่อเช้า

ลุง	ตู่	ต่อว่า	ผู้ สืบ ขอ ข่าว	ที่ ตึก ไทยคุ่ฟ้า	เมื่อ	เช้า
Noun	PNoun	Verb	Noun	Adj	PNoun	ADP
B-PER	I-PER	O	O	O	B-PLACE	B-TIME

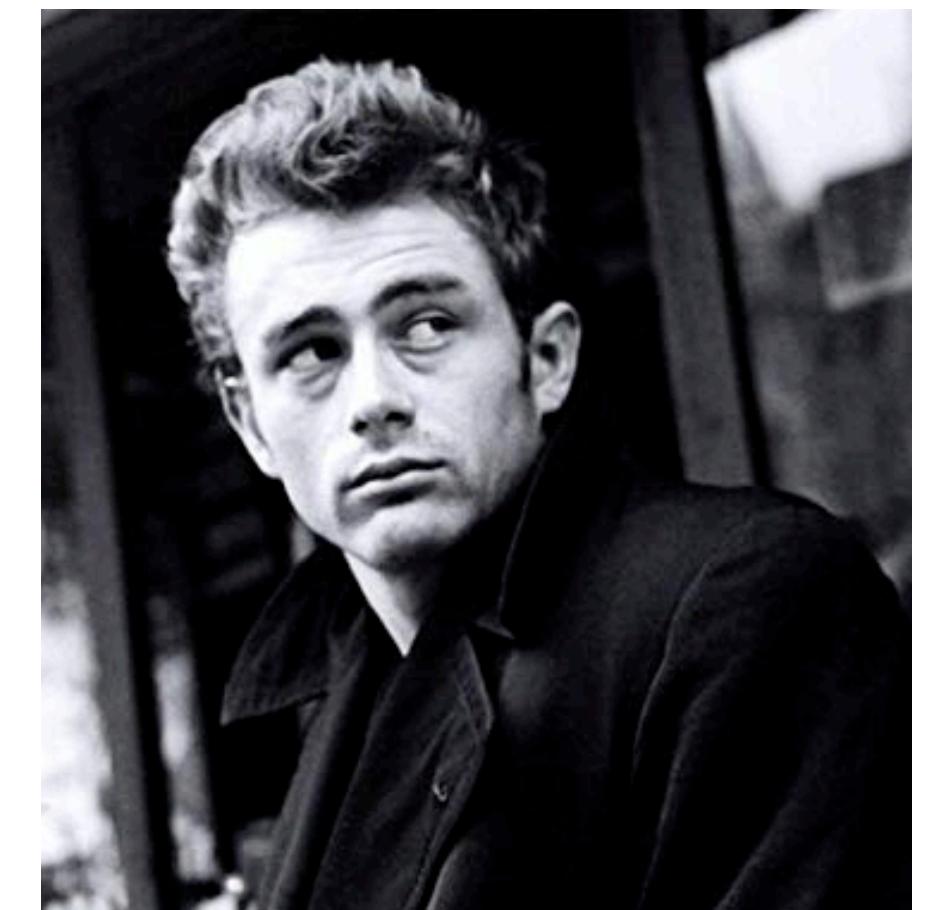
ชื่อคนไทย

- สนธยา คุณปลื้ม
- กุ้งอบวุ้นเส้น
- ปรัชญาเศรษฐกิจพอเพียง
- วัฒนาการะໂປຣງແດງ



ថ្មី ថ្មីទំនាំ

- Are you Rich?
- Play a song by Ke\$ha
- A new mission to Mars
- Dean's new single



ຊື່ອົນສໍ ຂໍອໂປຣຕິນ

- rpsL = ribosomal protein, small S12
- polA = DNA polymerase I
- gal = galactose
- cat = chloramphenicol resistance
- amp
- azi

ຂໍ້ອຍາ

- quetiapine = Seroquel XR
- PN = penicillin != pneumonia
- IUPAC = 7-{4-[4-(2,3-dichlorophenyl) piperazin-1-yl]
butoxy}-3,4-dihydroquinolin-2(1H)-one
- loop, potassium-sparing and thiazide diuretics

(Dai et al, 2017)

การรู้จำเอ็นทิชี

- NER มักถูกแก้ด้วย sequence labeling model โดยใช้ IOB label
- ยังจำเป็นต้องนิยามชัติของข้อมูลว่าประหลาดอย่างไร

Sequence Labeling Model

Sequence Labeling vs Classification

ลุง	ตู่	ต่อว่า	ผู้เสื่อข่าว	ที่	ติกไวยคุ่ฟ้า	เมื่อ	เช้า
Noun	PNoun	Verb	Noun	Adj	PNoun	ADP	Noun

Sequence ของหน่วยทางภาษาต่าง ๆ

ลุง	ตู่	ต่อว่า	ผู้สืบทอด	ที่	ตึกไทยคู่ฟ้า	เมื่อ	เช้า
B-PER	I-PER	O	O	O	B-PLACE	B-TIME	I-TIME
ล	ง	ต	บ	ต	อ	ว	า
BP	IP	IP	IP	IP	O	O	O

Sequence ของหน่วยทางภาษาต่างๆ

ଲ - ଏ ଟ ପ - ଟ ର - ର ବ - ବ ଗ - ଗ ଫ - ଫ ନ - ନ ଡ - ଡ ବ - ବ ଖ - ଖ ଶ - ଶ

B O O O O O B O O O O O B O

Sequence ของหน่วยทางภาษาต่าง ๆ

เรื่องนี้คนแสดงนำหล่อ แต่เนื้อเรื่องน่าเบื่อ จบได้จีดมาก

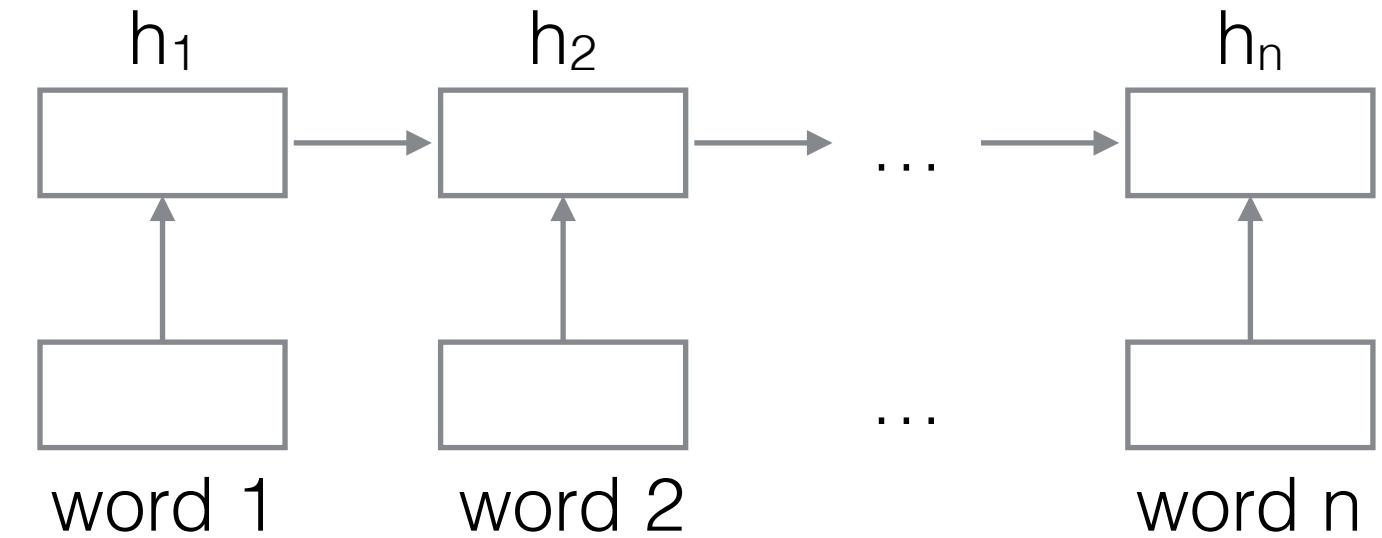
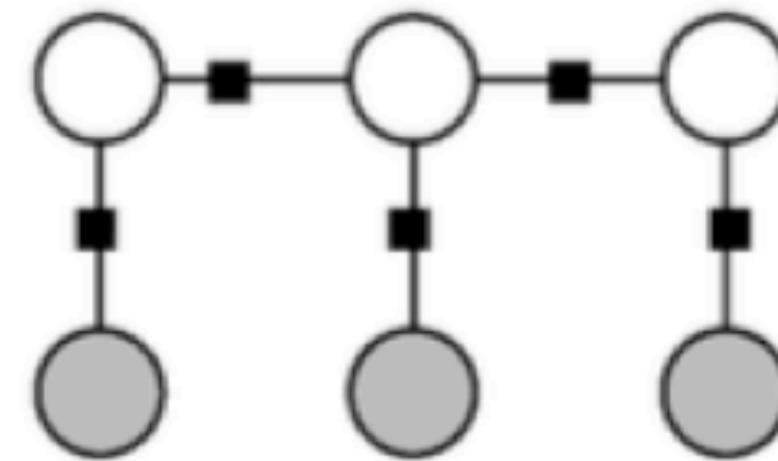
positive

negative

negative

Sequence Model ທີ່ອີຕອຍ່ຂະນະນີ້

- Conditional Random Fields (CRF)
- Recurrent Neural Network (RNN)
- RNN + CRF



Sequence Labeling Model

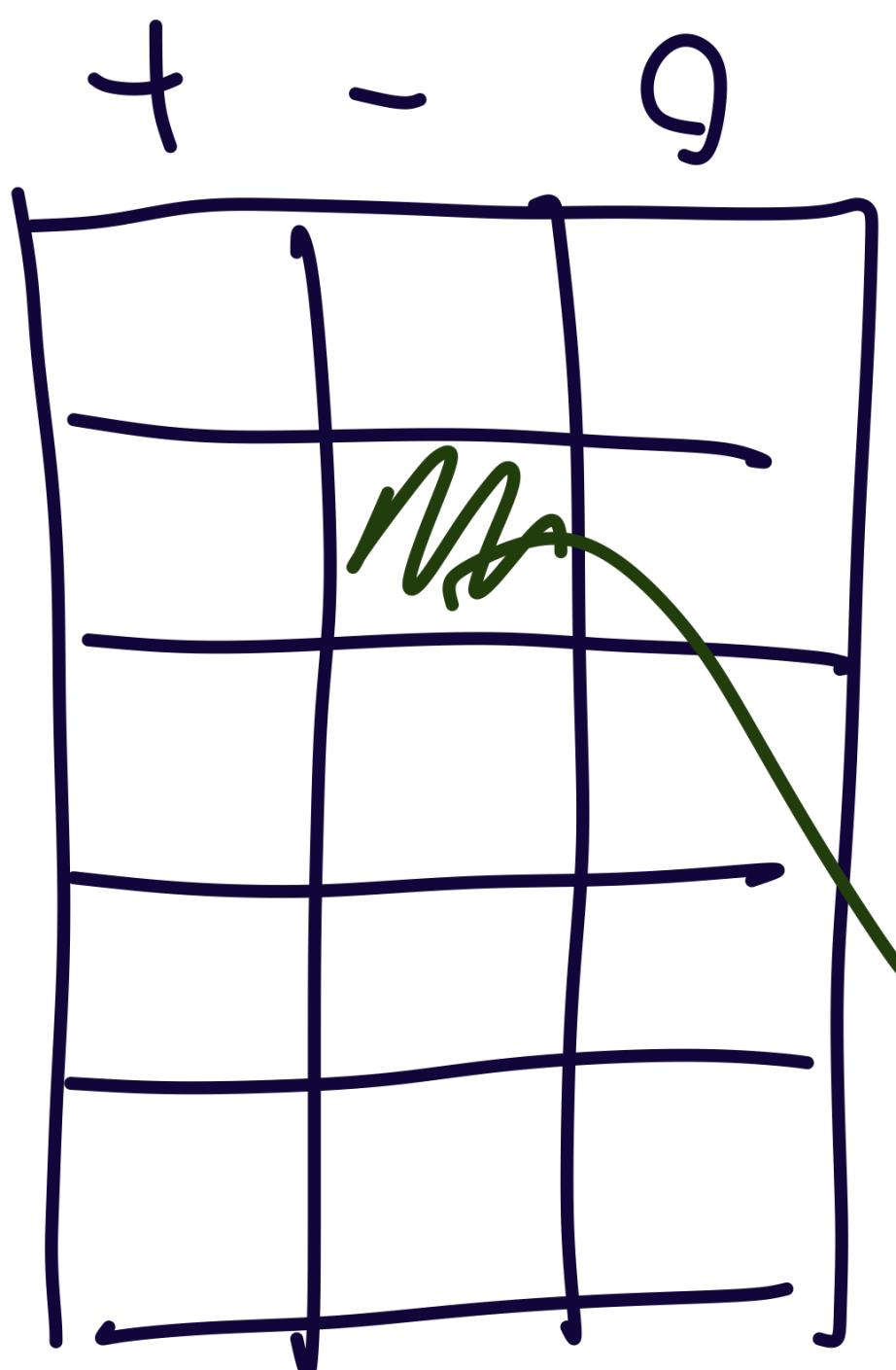
- ภาษา มีการเรียงตัวกันเป็นลำดับ
- ผลจะแม่นยำขึ้นถ้า Label มีความเกี่ยวเนื่องกันใน sequence
และจำนวน Label = จำนวนหน่วย

Conditional Random Fields (CRF)

MaxEnt



bias

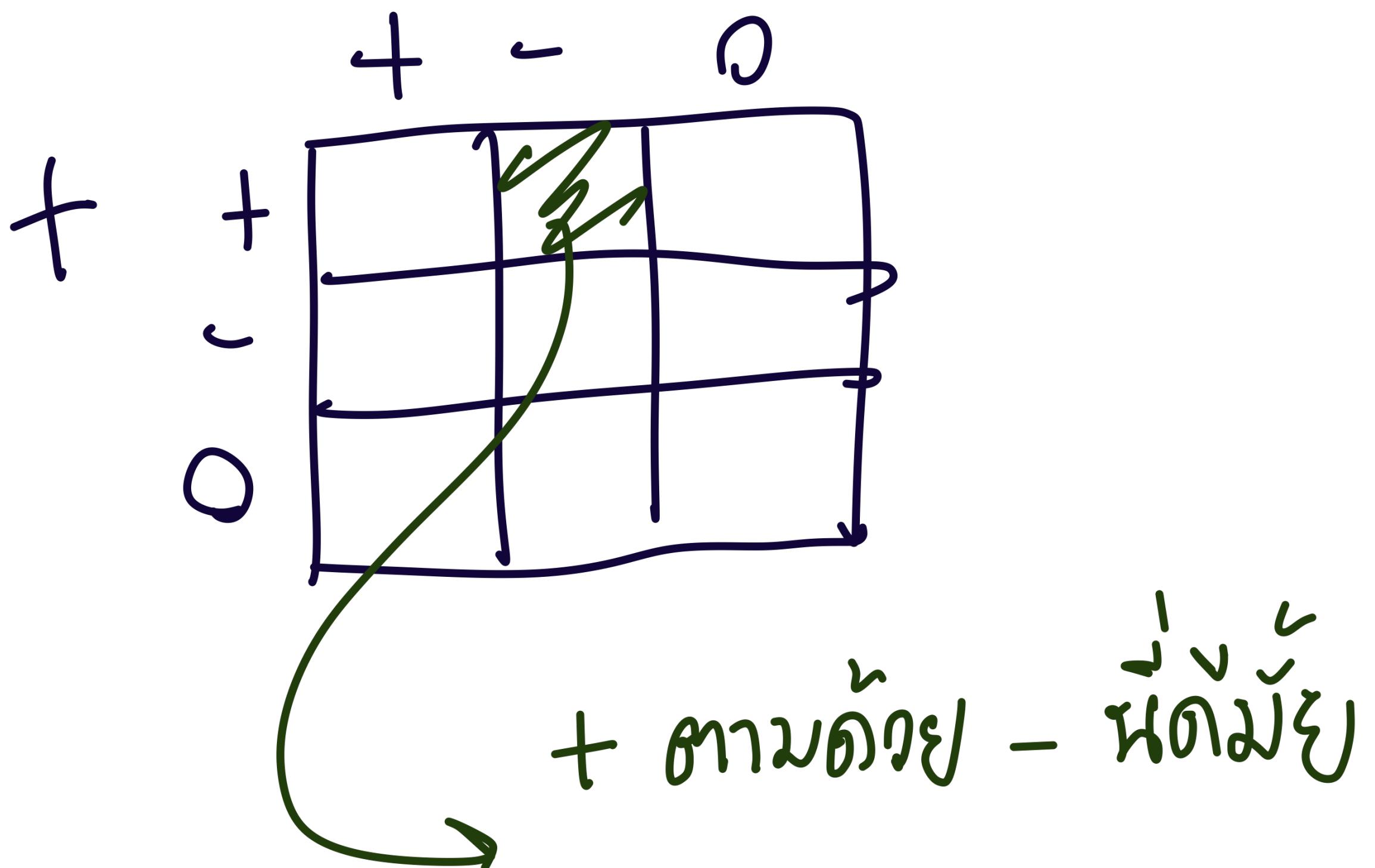


parameter
matrix

feature
กับ - หัวมันๆ

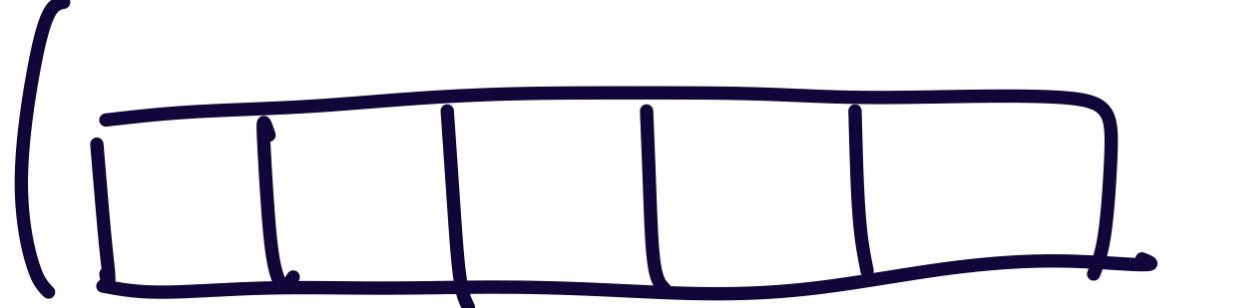
CRF

label compatibility

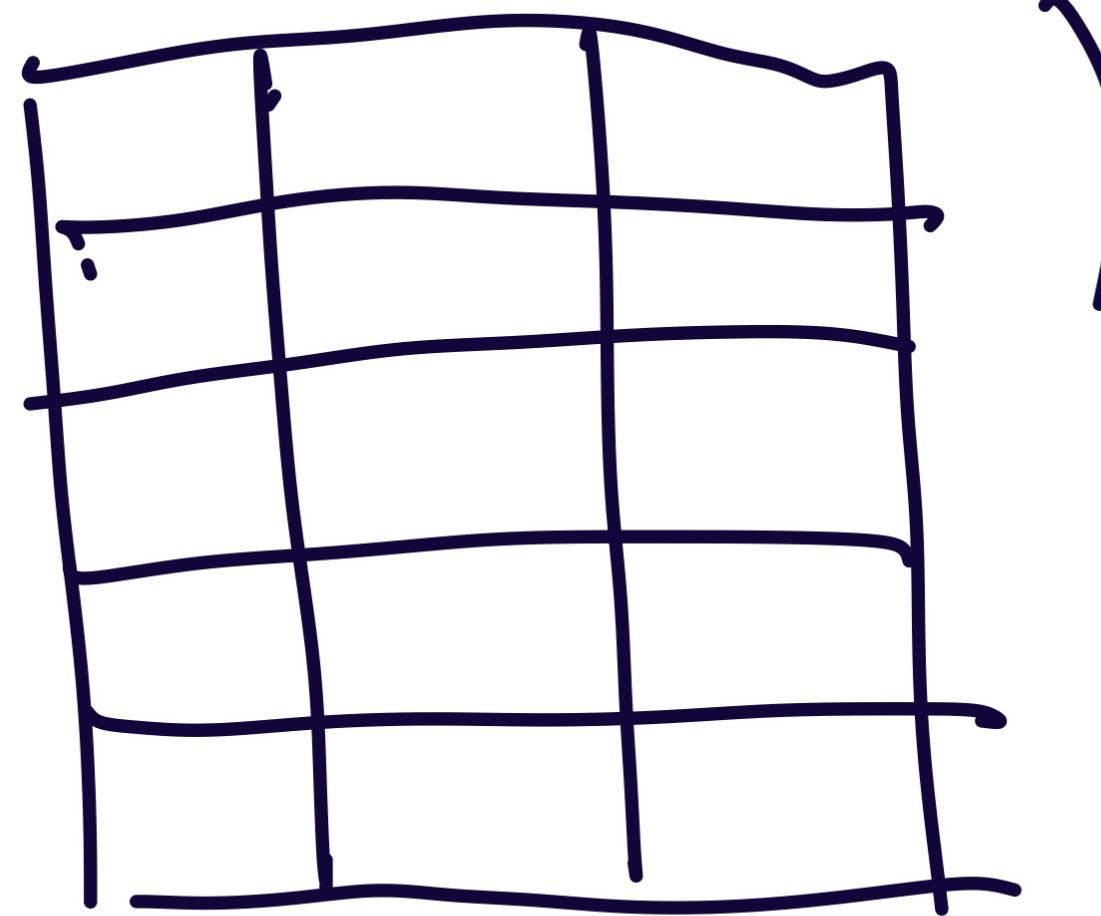


+ ตามด้วย - หัวมันๆ

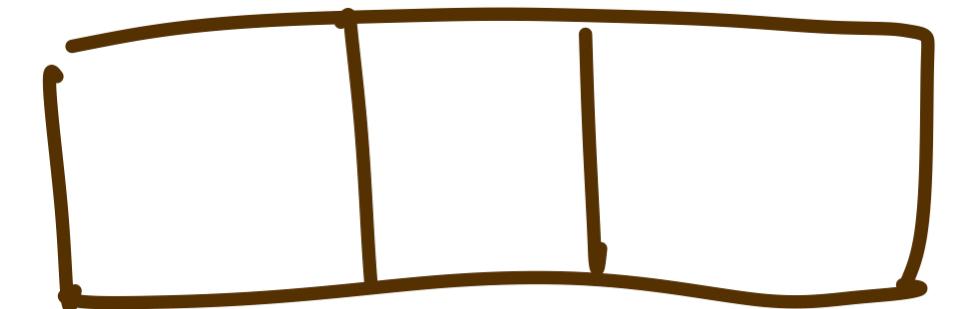
feature vector



parameter matrix



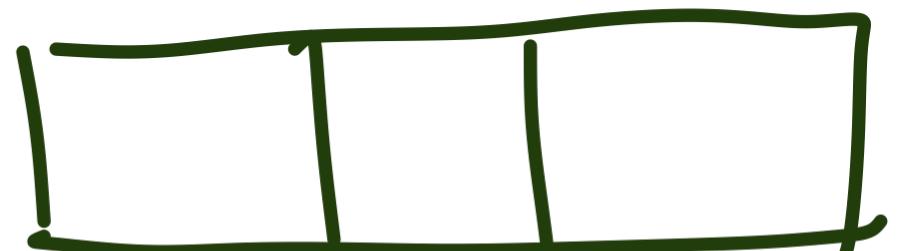
bias



+ - 0

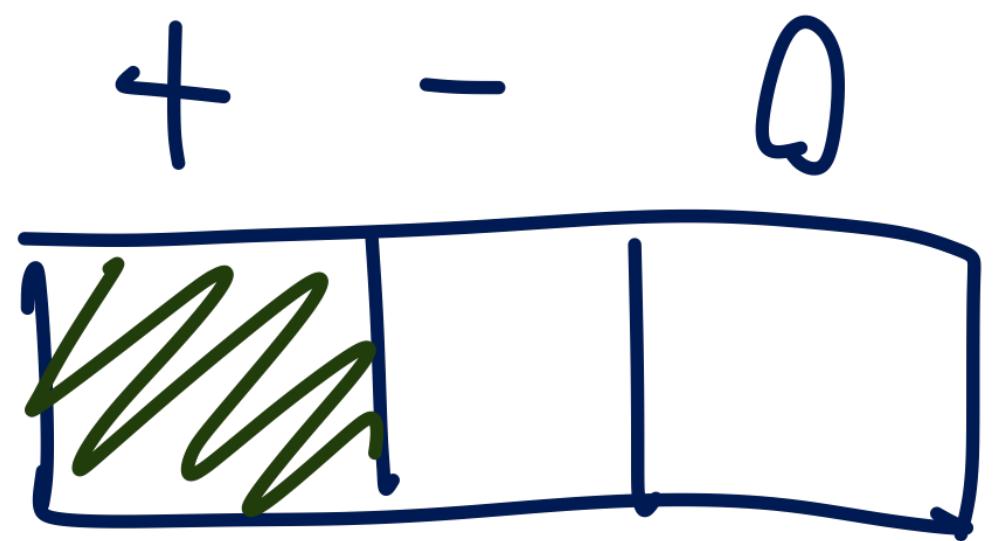
$k = 3$

unnormalized score

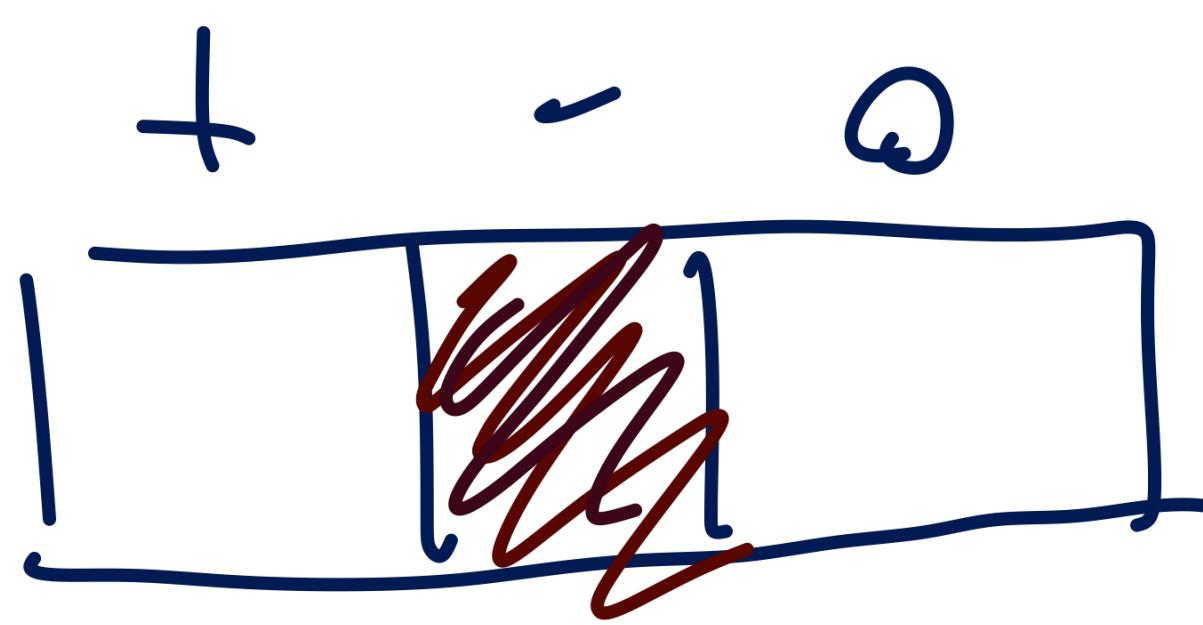


+ - 0

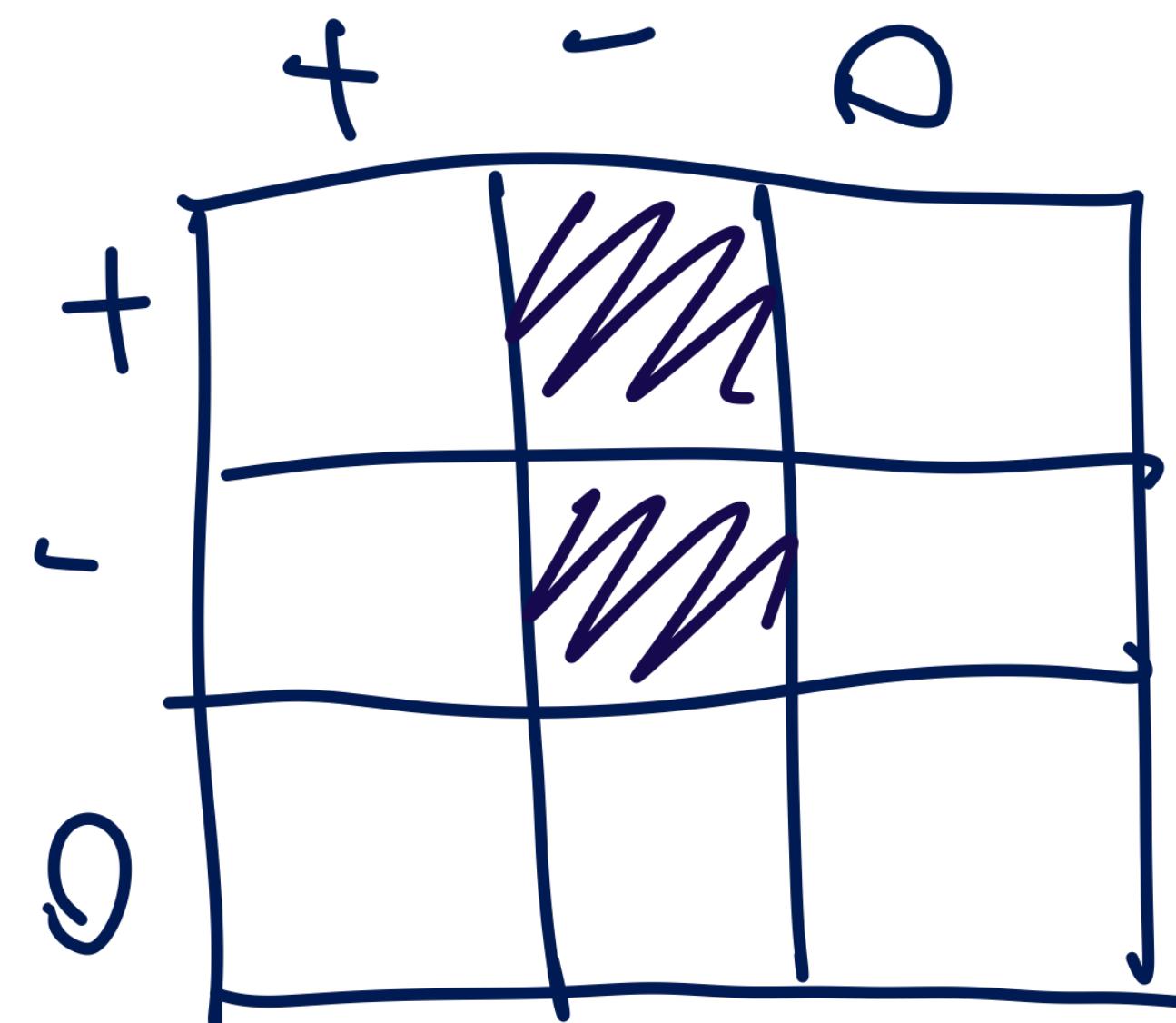
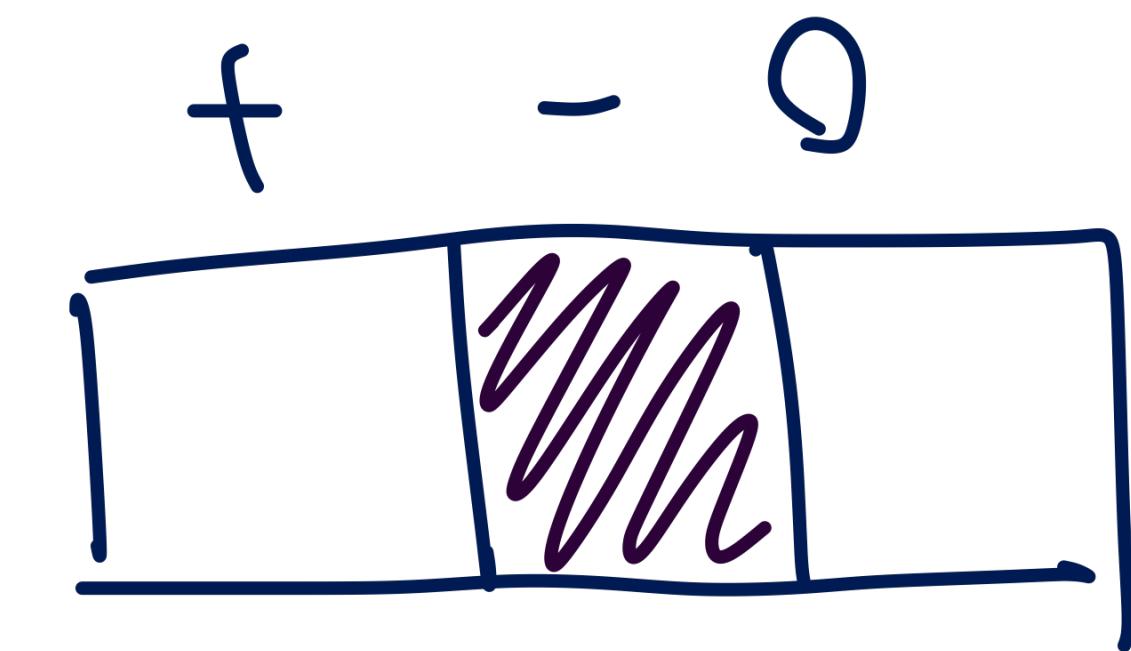
$t=0$
sentence



$t=1$
sentehce



$t=2$
sentence



$$\text{score}(+, -, -) = \text{Score}_0(+) + \text{Score}_1(-) + \text{Score}_2(-) + \text{tScore}(+, -) + \text{tScore}(-, -)$$

$3 \times 3 \times 3$

combinations = too slow

label sequence	unnormalized	Probability
+++		
++-		
++0		
+ - +		
+ - -		
+ - 0		
-		
;		

$$\text{sum} = 1$$

Conditional Random Fields

- Training ต้องใช้ algorithm ที่หา probability ได้เร็วๆ
- Decoding ต้องใช้ algorithm ที่หา label sequence ที่ดีสุดได้เร็วๆ

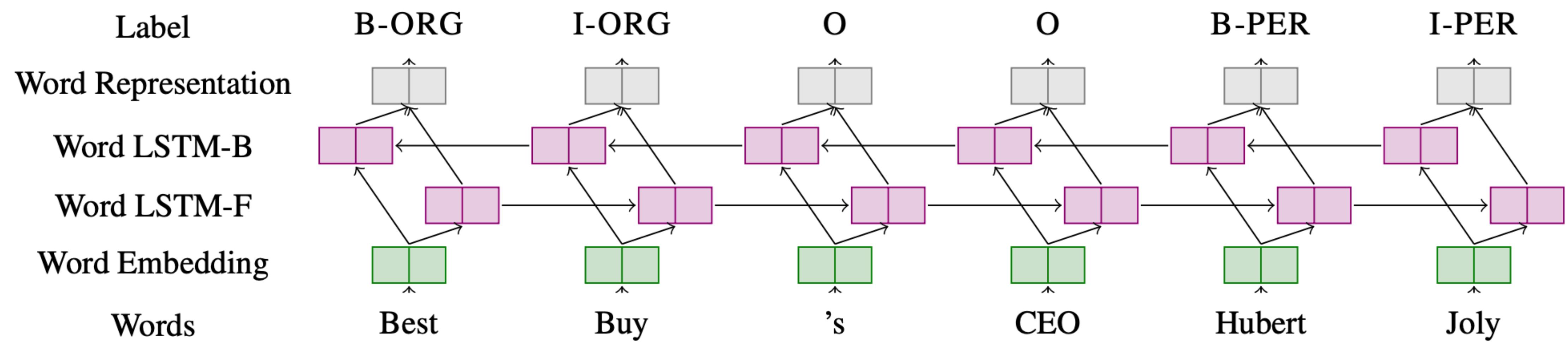
Features for Named Entity Recognition

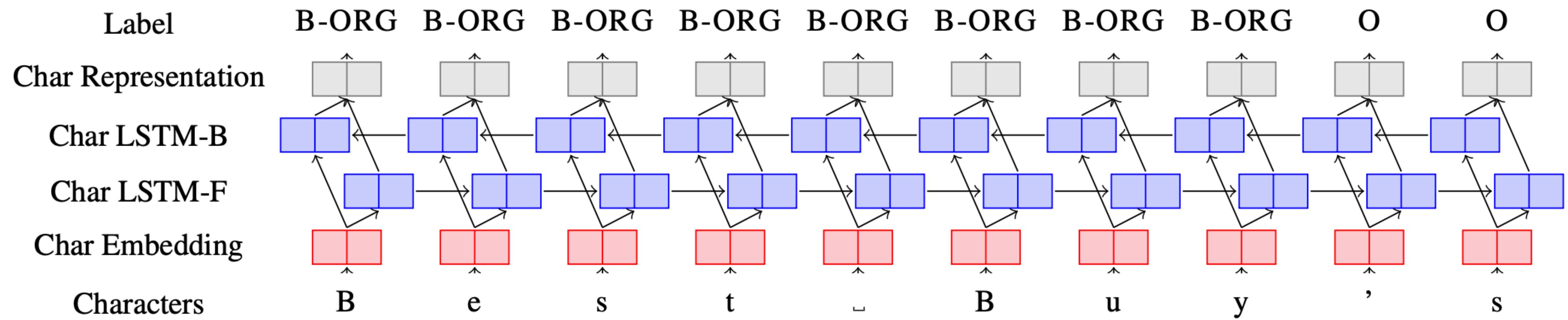
Features ที่ใช้มีผลมาก ๆ

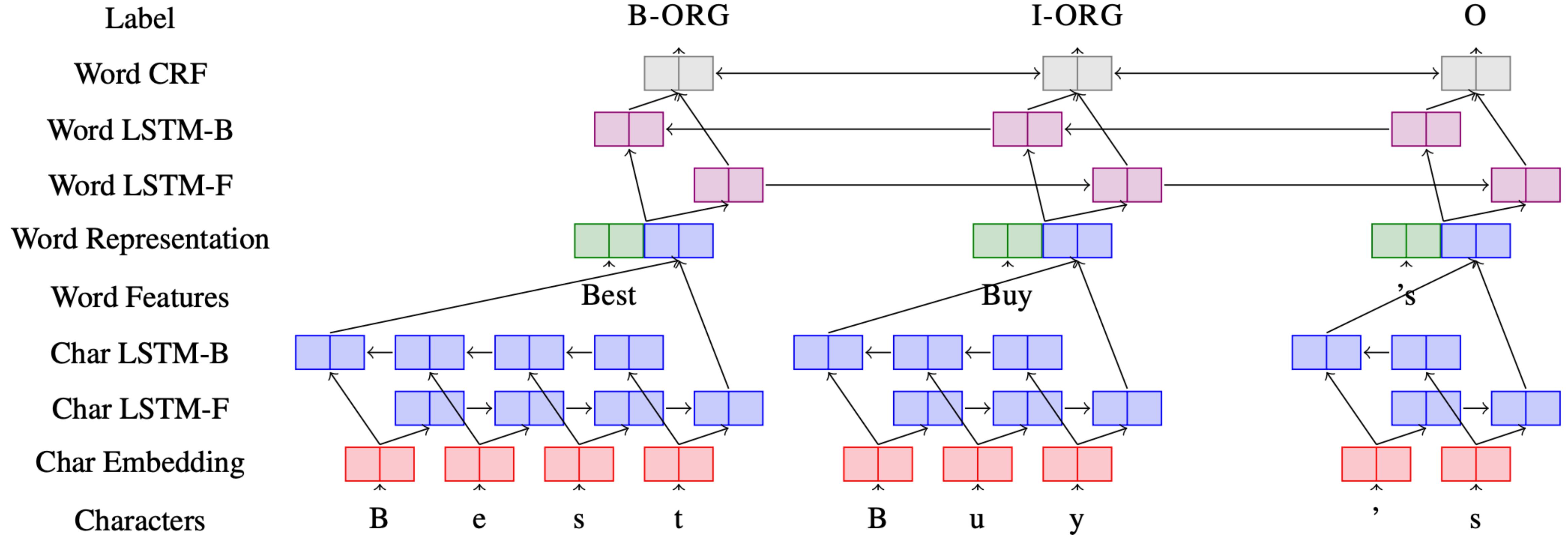
- Character-level features e.g. word shape, prefix, suffix
- Word-level features
- Word type features
- World knowledge features

Bruno Mars goes to Mars

Neural Network for NER







Feature-engineered machine learning systems	Dict	SP	DU	EN	GE
Carreras et al. (2002) binary AdaBoost classifiers	Yes	81.39	77.05	-	-
Malouf (2002) - Maximum Entropy (ME) + features	Yes	73.66	68.08	-	-
Li et al. (2005) SVM with class weights	Yes	-	-	88.3	-
Passos et al. (2014) CRF	Yes	-	-	90.90	-
Ando and Zhang (2005a) Semi-supervised state of the art	No	-	-	89.31	75.27
Agerri and Rigau (2016)	Yes	84.16	85.04	91.36	76.42
Feature-inferring neural network word models					
Collobert et al. (2011) Vanilla NN +SLL / Conv-CRF	No	-	-	81.47	-
Huang et al. (2015) Bi-LSTM+CRF	No	-	-	84.26	-
Yan et al. (2016) Win-BiLSTM (English), FF (German) (Many fets)	Yes	-	-	88.91	76.12
Collobert et al. (2011) Conv-CRF (SENNA+Gazetteer)	Yes	-	-	89.59	-
Huang et al. (2015) Bi-LSTM+CRF+ (SENNA+Gazetteer)	Yes	-	-	90.10	-
Feature-inferring neural network character models					
Gillick et al. (2015) – BTS	No	82.95	82.84	86.50	76.22
Kuru et al. (2016) CharNER	No	82.18	79.36	84.52	70.12
Feature-inferring neural network word + character models					
Yang et al. (2017)	Yes	85.77	85.19	91.26	-
Luo (2015)	Yes	-	-	91.20	-
Chiu and Nichols (2015)	Yes	-	-	91.62	-
Ma and Hovy (2016)	No	-	-	91.21	-
Santos and Guimaraes (2015)	No	82.21	-	-	-
Lample et al. (2016)	No	85.75	81.74	90.94	78.76
Bharadwaj et al. (2016)	Yes	85.81	-	-	-
Dernoncourt et al. (2017)	No	-	-	90.5	-
Feature-inferring neural network word + character + affix models					
Re-implementation of Lample et al. (2016) (100 Epochs)	No	85.34	85.27	90.24	78.44
Yadav et al. (2018)(100 Epochs)	No	86.92	87.50	90.69	78.56
Yadav et al. (2018) (150 Epochs)	No	87.26	87.54	90.86	79.01

	Dict	MedLine (80.10%)			DrugBank (19.90%)			Complete dataset		
		P	R	F1	P	R	F1	P	R	F1
Feature-engineered machine learning systems										
Rocktäschel et al. (2013)	Yes	60.70	55.80	58.10	88.10	87.50	87.80	73.40	69.80	71.50
Liu et al. (2015) (baseline)	No	-	-	-	-	-	-	78.41	67.78	72.71
Liu et al. (2015) (MED. emb.)	No	-	-	-	-	-	-	82.70	69.68	75.63
Liu et al. (2015) (state of the art)	Yes	78.77	60.21	68.25	90.60	88.82	89.70	84.75	72.89	78.37
NN word model										
Chalapathy et al. (2016) (relaxed performance)	No	52.93	52.57	52.75	87.07	83.39	85.19	-	-	-
NN word + character model										
Yadav et al. (2018)	No	73	62	67	87	86	87	79	72	75
NN word + character + affix model										
Yadav et al. (2018)	No	74	64	69	89	86	87	81	74	77

Training CRF

Simple Classification

label	sentence
+	S_1
-	S_2
-	S_3
0	S_4
-	S_5
+	S_6
+	S_7
+	S_8

Sequence Labeling

label sequence	sentence sequence
$+, -, -$	S_1, S_2, S_3
$0, -$	S_4, S_5
$+, +, +$	S_6, S_7, S_8

$3 \times 3 \times 3$

combinations = too slow

label sequence	unnormalized	Probability
+++		
++-		
++0		
+ - +		
+ - -		
+ - 0		
-		
;		

$$\text{sum} = 1$$

Objective Function

$$L(\theta) = - \sum_{\text{Sequences } \bar{y} \in D} \log P(\bar{y} | \bar{x})$$

sequence of feature vectors
sequence of labels

Training Algorithm

- Forward-Backward algorithm
- Averaged Structured Perceptron

Forward-Backward Algorithm

- คำนวณ log-likelihood ของ data (forward)
- คำนวณ gradient ของแต่ละ parameter (forward-backward)

Averaged Structured Perceptron

- Viterbi algorithm ในการ decode แล้วปรับแก้ parameter โดยไม่ต้องใช้ gradient

Training

`class pycrfsuite.Trainer`

Bases: `pycrfsuite._pycrfsuite.BaseTrainer`

The trainer class.

This class maintains a data set for training, and provides an interface to various training algorithms.

Parameters: `algorithm:{'lbfsgs', 'l2sgd', 'ap', 'pa', 'arow'}`

The name of the training algorithm. See `Trainer.select()`.

`params:dict, optional`

Training parameters. See `Trainer.set_params()` and `Trainer.set()`.

`verbose:boolean`

Whether to print debug messages during training. Default is True.

The name of the training algorithm.

- ‘lbfsgs’ for Gradient descent using the L-BFGS method,
- ‘l2sgd’ for Stochastic Gradient Descent with L2 regularization term
- ‘ap’ for Averaged Perceptron
- ‘pa’ for Passive Aggressive
- ‘arow’ for Adaptive Regularization Of Weight Vector

Workshop วันนี้

- <https://attapol.github.io/talks>