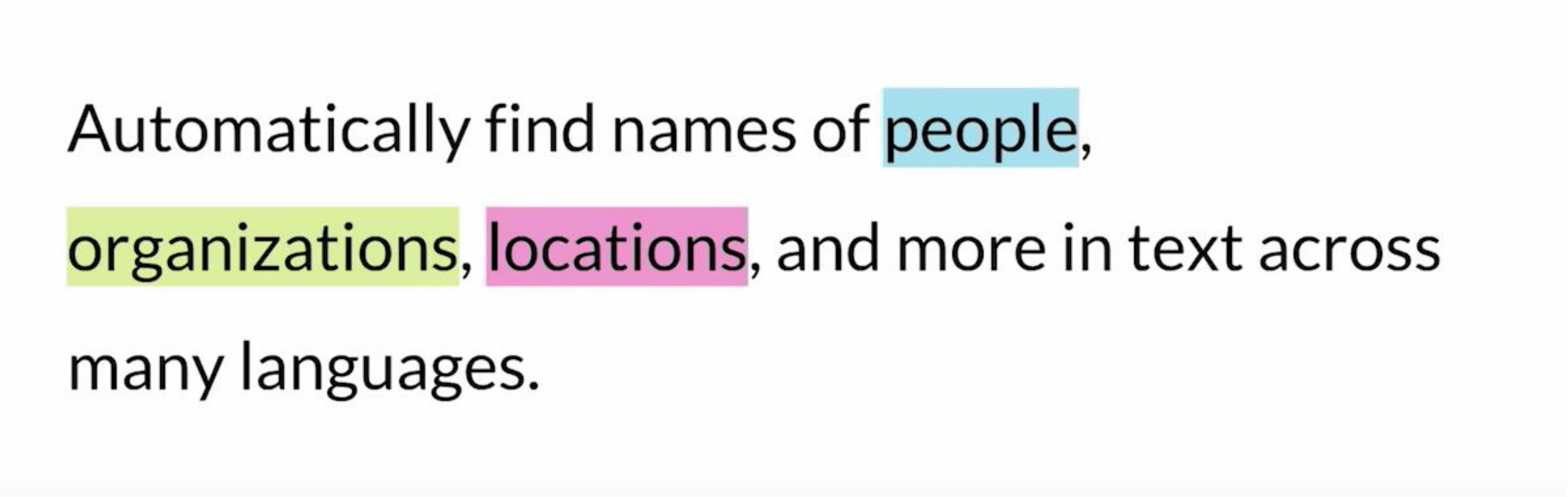
Attribute Value Extraction from Product Profiles

Named entity recognition (NER) in business project

Outline

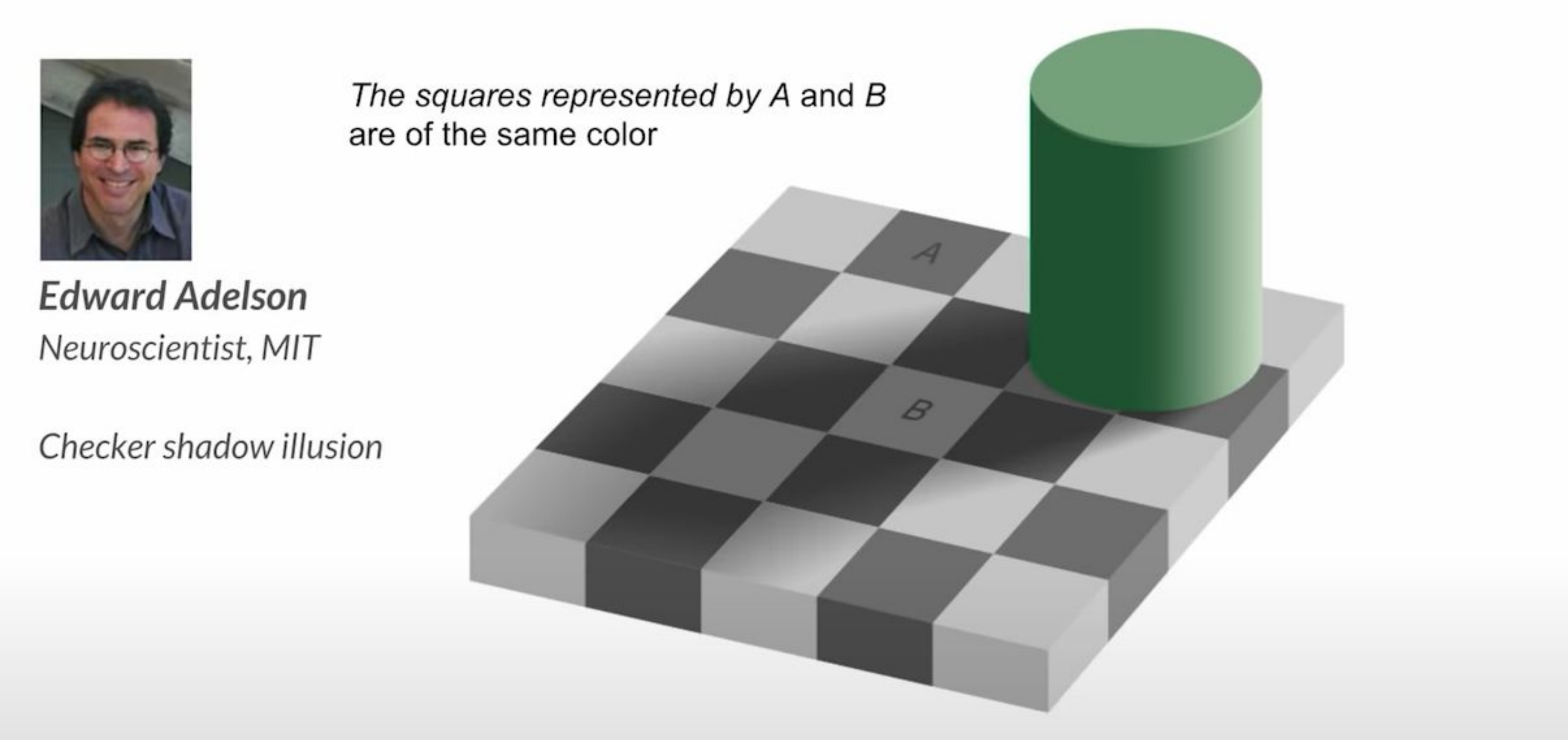
- NER Models survey
- Product attributes tagging
- Challenges and low-resource training

Named Entity Recognition (NER)



Automatically find names of people, organizations, locations, and more in text across many languages.

Context is important in NLP problems



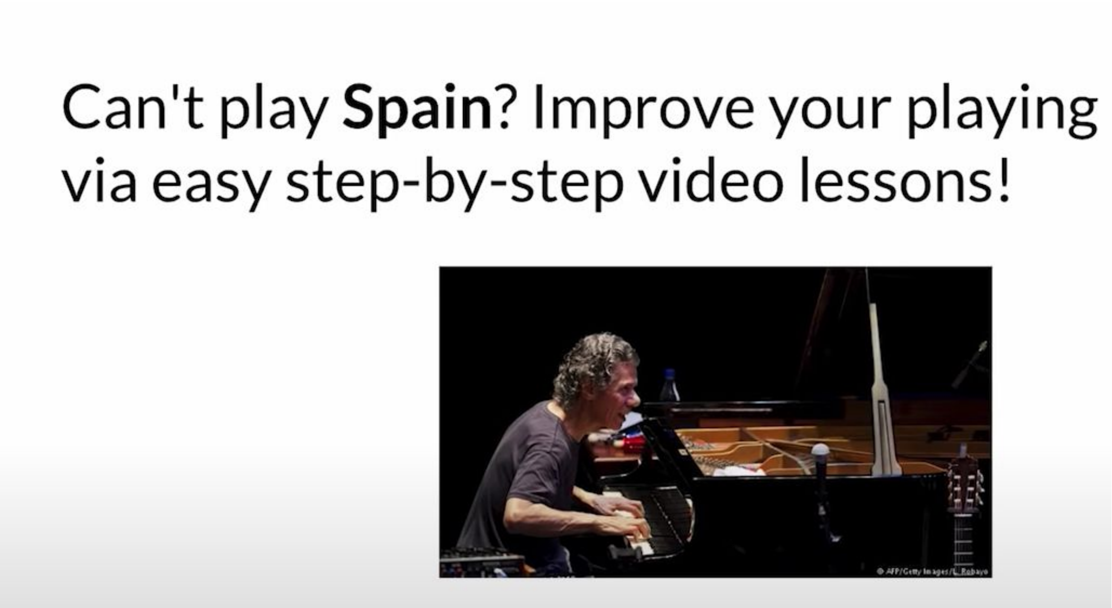
The squares represented by A and B are of the same color

Edward Adelson

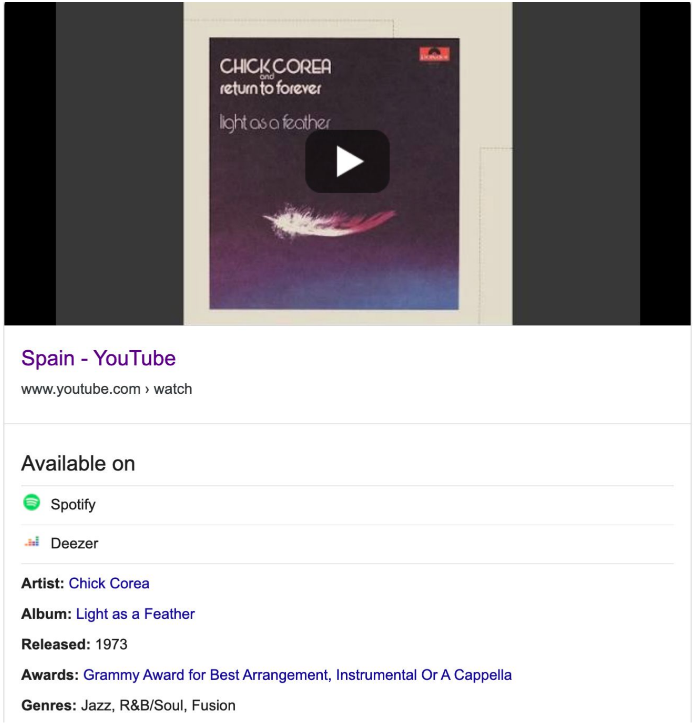
Neuroscientist, MIT

Checker shadow illusion

But sometimes it gets ambiguous...



Can't play **Spain**? Improve your playing via easy step-by-step video lessons!



Spain - YouTube

www.youtube.com › watch

Available on

Spotify

Deezer

Artist: Chick Corea

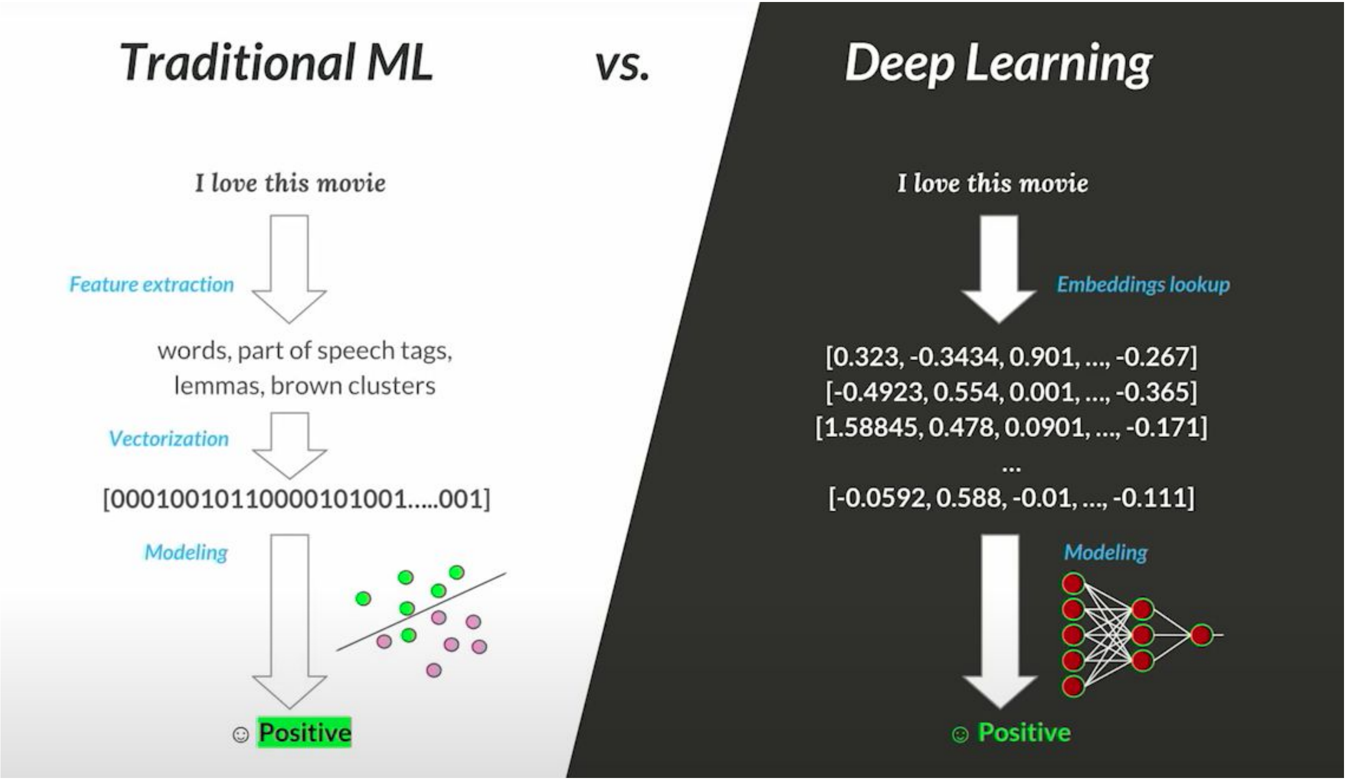
Album: Light as a Feather

Released: 1973

Awards: Grammy Award for Best Arrangement, Instrumental Or A Cappella

Genres: Jazz, R&B/Soul, Fusion

Why all in deep learning ?



Traditional ML vs. Deep Learning

I love this movie

Feature extraction

words, part of speech tags, lemmas, brown clusters

Vectorization

[00010010110000101001.....001]

Modeling

☺ Positive

I love this movie

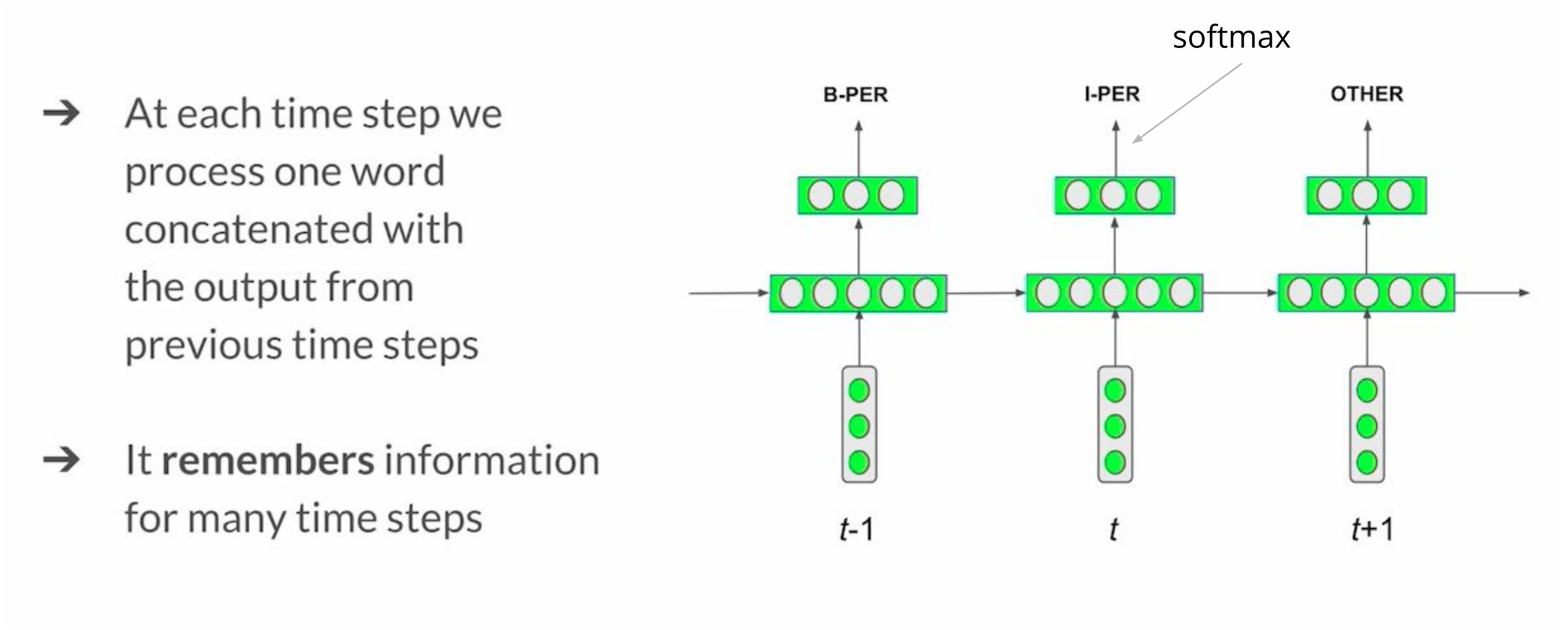
Embeddings lookup

[0.323, -0.3434, 0.901, ..., -0.267]
[-0.4923, 0.554, 0.001, ..., -0.365]
[1.58845, 0.478, 0.0901, ..., -0.171]
...
[-0.0592, 0.588, -0.01, ..., -0.111]

Modeling

☺ Positive

Recurrent neural network (RNN)



- ➔ At each time step we process one word concatenated with the output from previous time steps
- ➔ It **remembers** information for many time steps

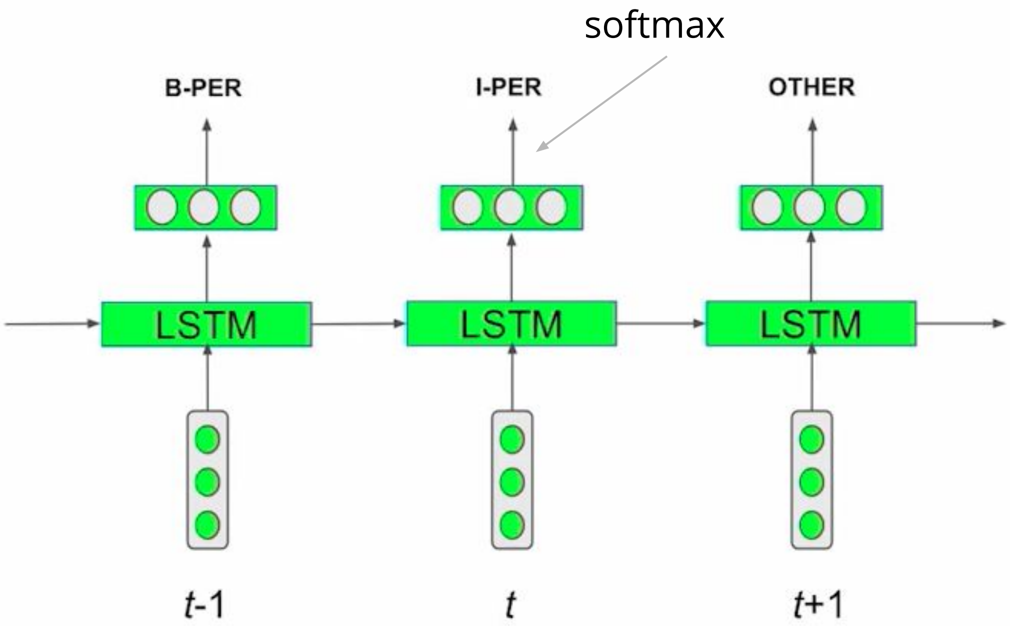
softmax

Long Short Term Memory (LSTM)

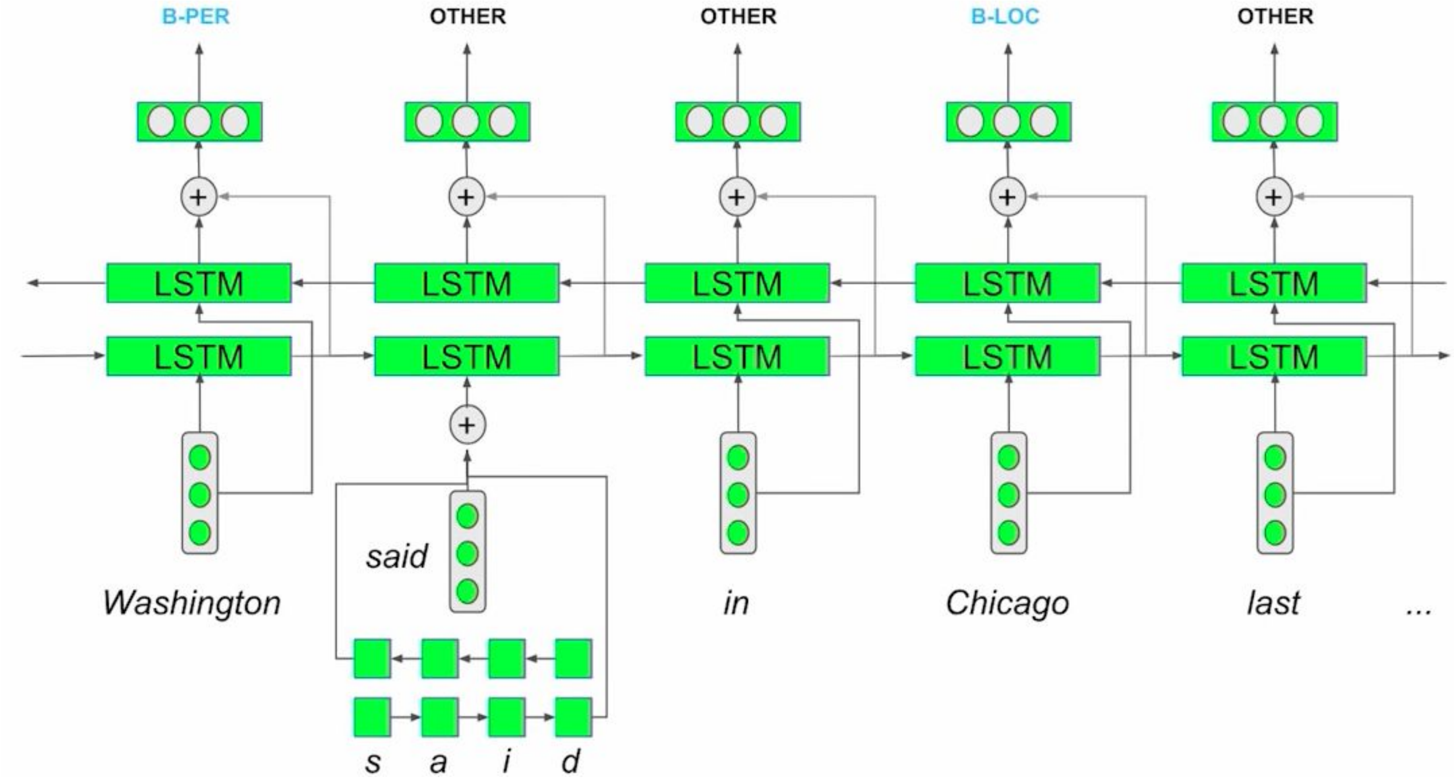
It can **forget** information when necessary

Prevent from

Gradient vanishing &
Gradient explosion

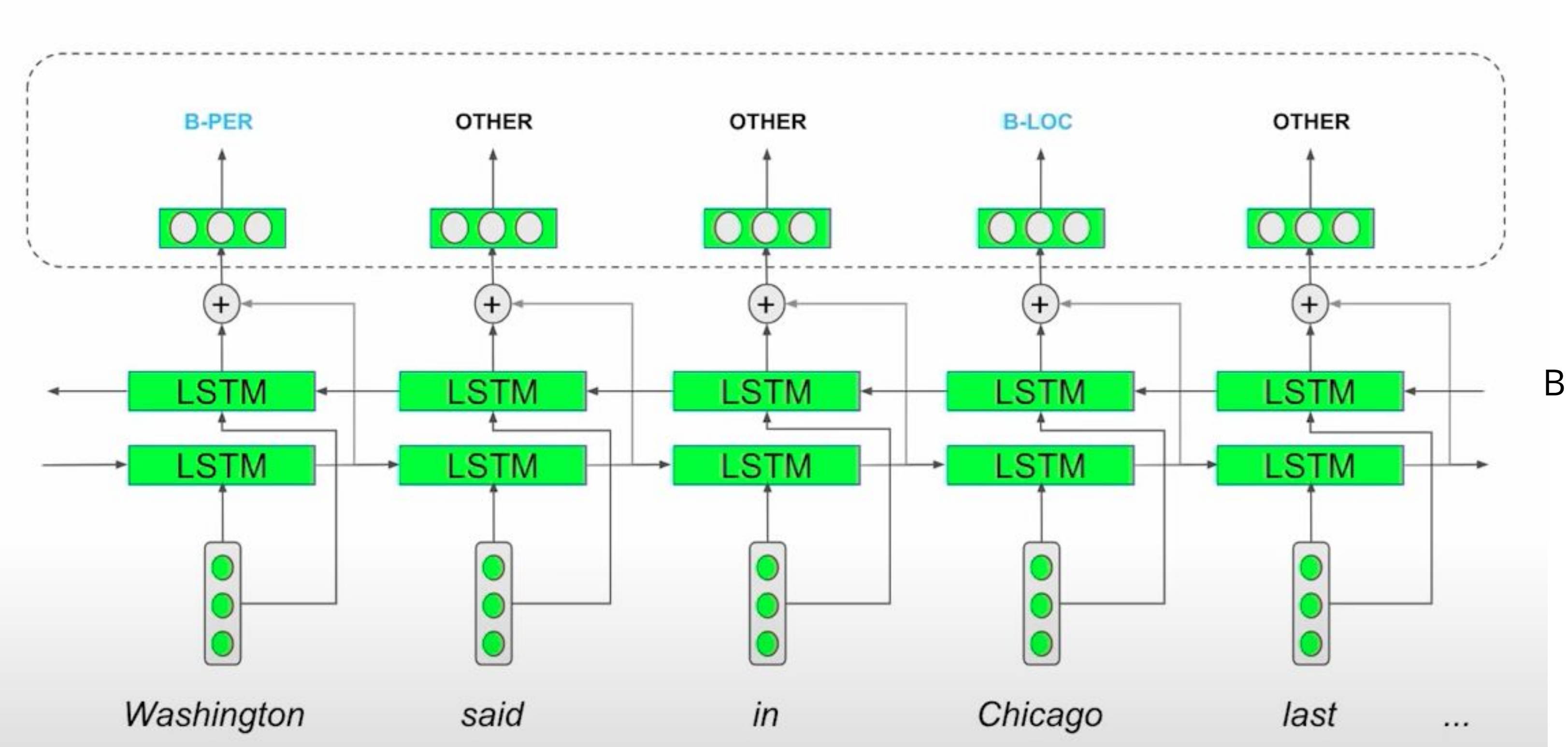


Character encoding



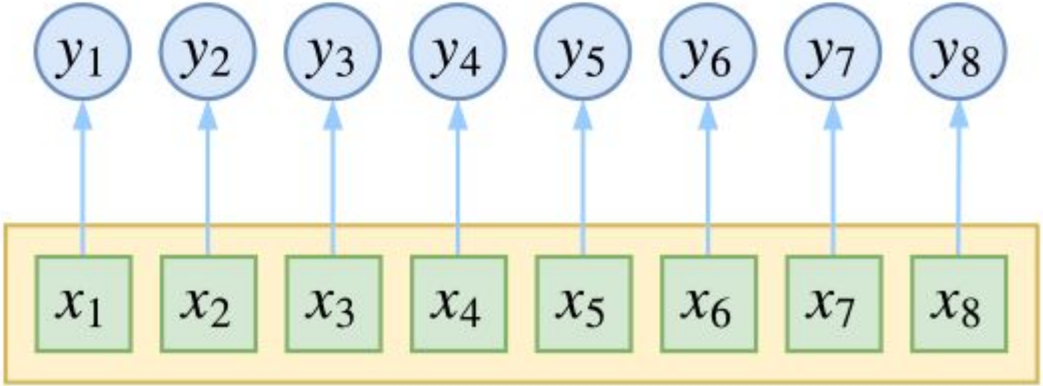
Bi-LSTM

Alternative decoding using Conditional Random Fields (CRF)

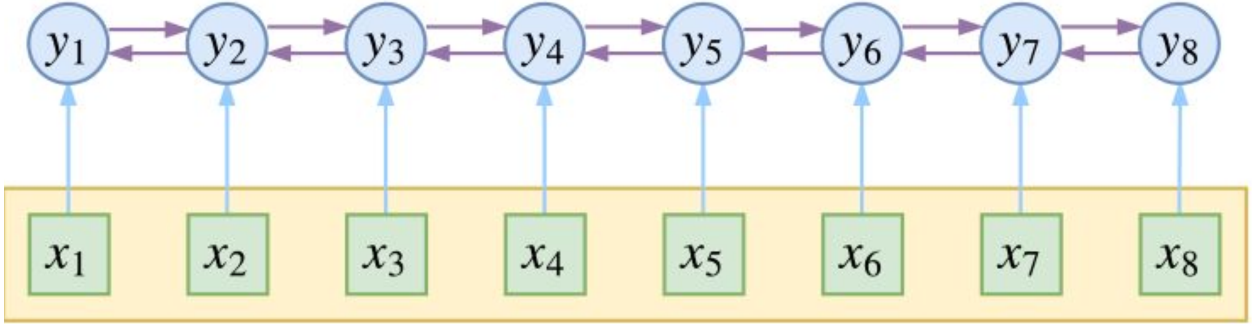


Bi-LSTM

CRF is different with softmax


y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8

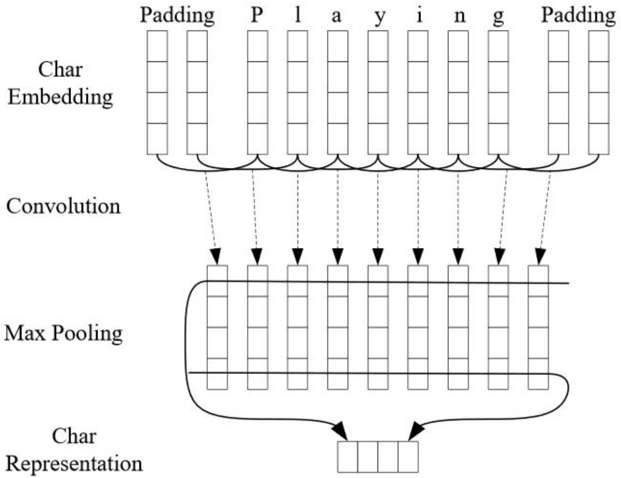
softmax:
N K-Classification


y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8

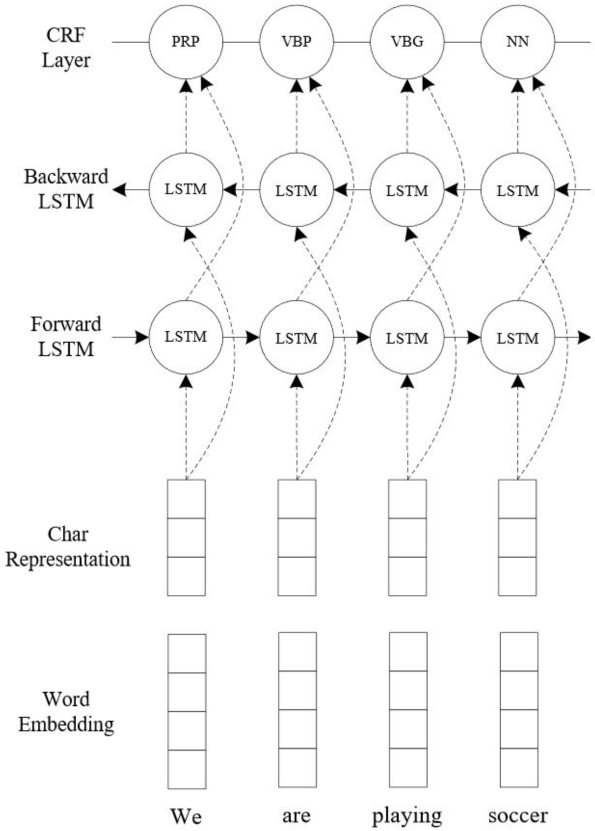
CRF:

K^N possible candidates
and select the best one

Bi-directional LSTM-CNNs-CRF

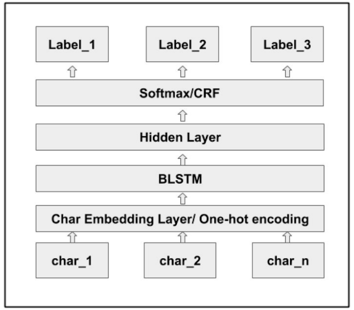


(a) CNN to obtain character-level representations. Dotted lines indicate dropout.

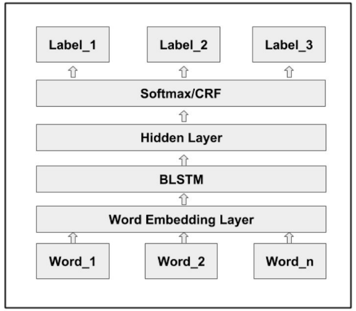


(b) BiLSTM network with CRF decoding layer. At the input, word embedding is concatenated with character representation.

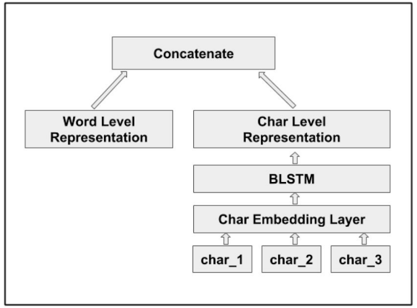
- Ma and Hovy achieve state-of-the-art F1 score of 91.21 for NER on CoNLL 2003 dataset.
- No feature engineering or specific data pre-processing.



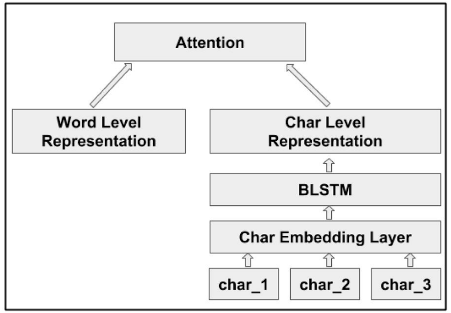
(**a**) CharModel



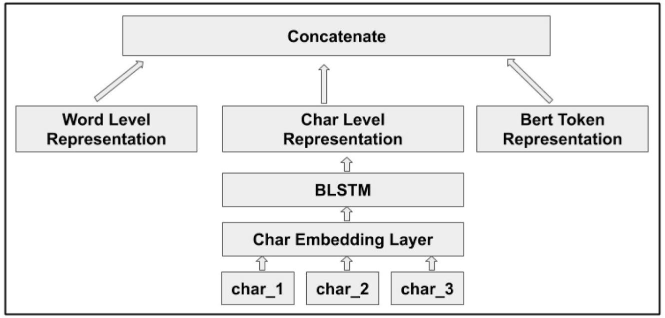
(**b**) WordModel



(**c**) WordCharacter



(**d**) WordCharacterAttention


Concatenate
Word Level Representation
Char Level Representation
Bert Token Representation
BLSTM
Char Embedding Layer
char_1
char_2
char_3

(**e**) WordCharacterBERT

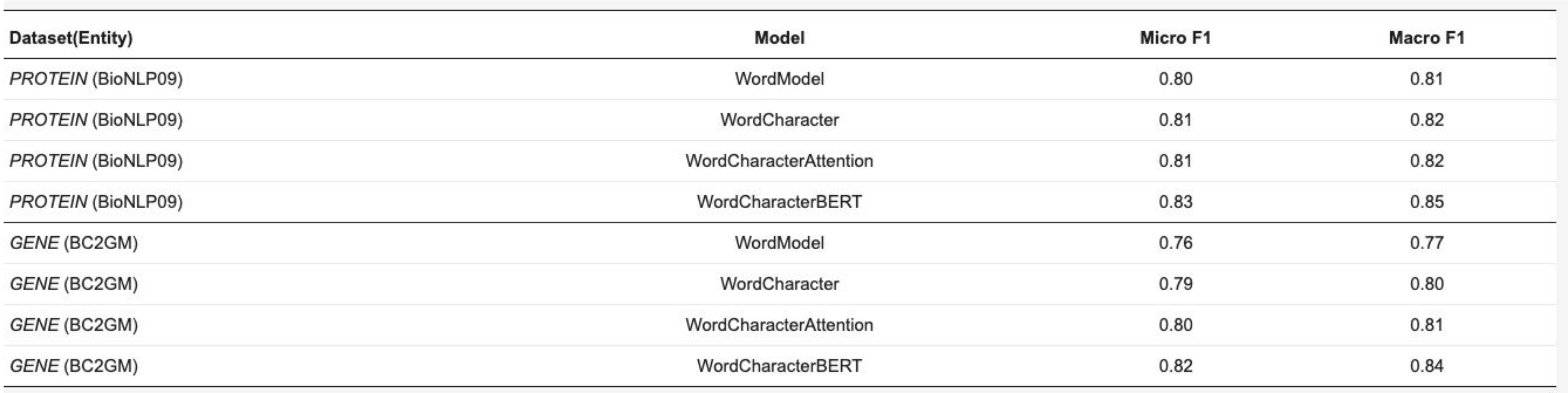
$$z = \sigma\left(W_z^{(3)}\ tanh\left(W_z^{(1)}x + W_z^{(2)}k\right)\right) \tag{14}$$

where $W^{(1)}$, $W^{(2)}$ and $W^{(3)}$ are weight matrices for calculating z and σ is a sigmoid logistic function. z is the weight matrix between word representation x and character representation k.

$$x_{att} = z \cdot x + (1 - z) \cdot k \tag{15}$$

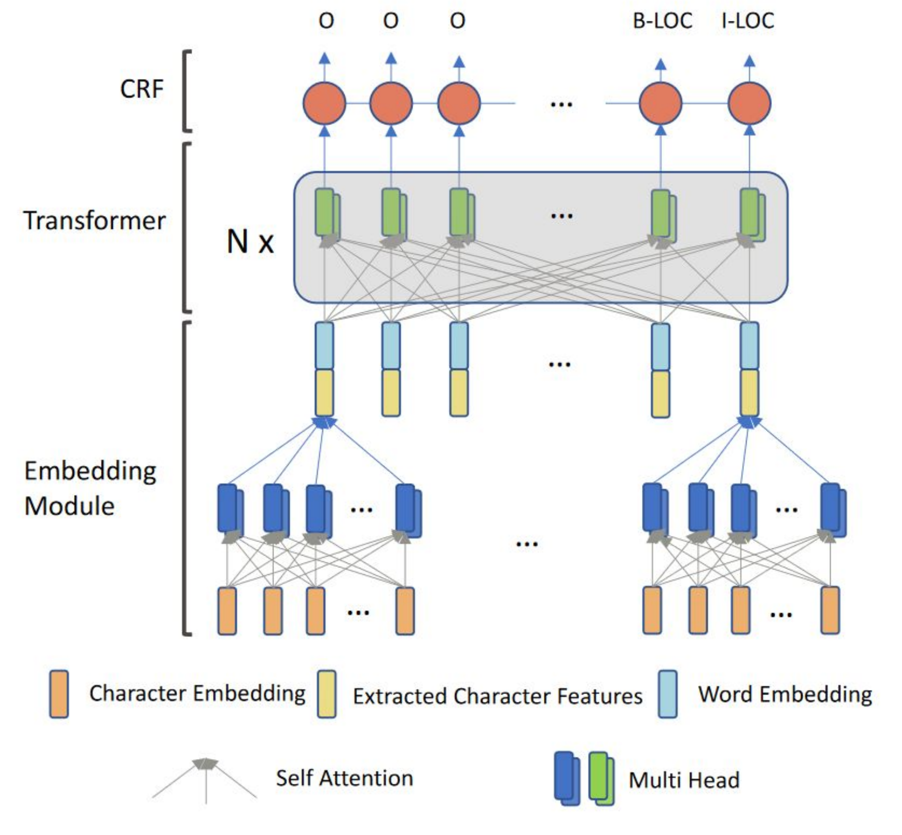
BERT can get the best performance

Micro- and macro-averaged F1 scores for NER in BioNLP09 and BC2GM.



Dataset(Entity)	Model	Micro F1	Macro F1
PROTEIN (BioNLP09)	WordModel	0.80	0.81
PROTEIN (BioNLP09)	WordCharacter	0.81	0.82
PROTEIN (BioNLP09)	WordCharacterAttention	0.81	0.82
PROTEIN (BioNLP09)	WordCharacterBERT	0.83	0.85
GENE (BC2GM)	WordModel	0.76	0.77
GENE (BC2GM)	WordCharacter	0.79	0.80
GENE (BC2GM)	WordCharacterAttention	0.80	0.81
GENE (BC2GM)	WordCharacterBERT	0.82	0.84

NER with transformer, first successful version



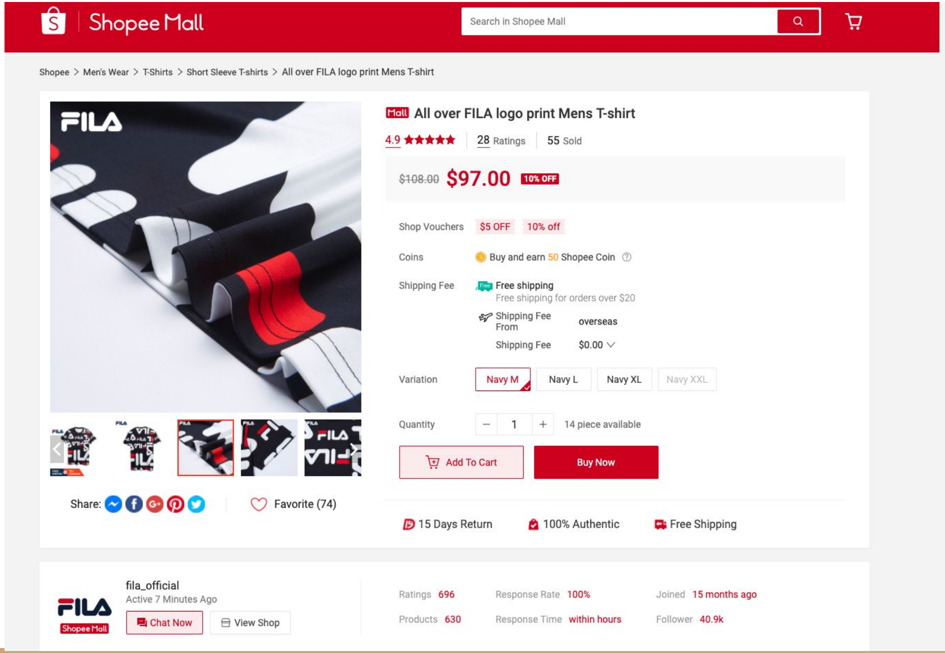
Customize transformer architecture and achieve state-of-the-art results, beating prevailing BiLSTM models.

2 Product attribute tagging

- 2.1 Background and problem definition
- 2.2 Data collection
- 2.3 Baseline model
- 2.4 Pain points
- 2.5 Conclusions and future work

2.1 Background and problem definition

Background: In Shopee, Many sellers do not fill in attributes, especially for some big sellers.



Shopee > Men's Wear > T-Shirts > Short Sleeve T-shirts > All over FILA logo print Mens T-shirt

Mall All over FILA logo print Mens T-shirt

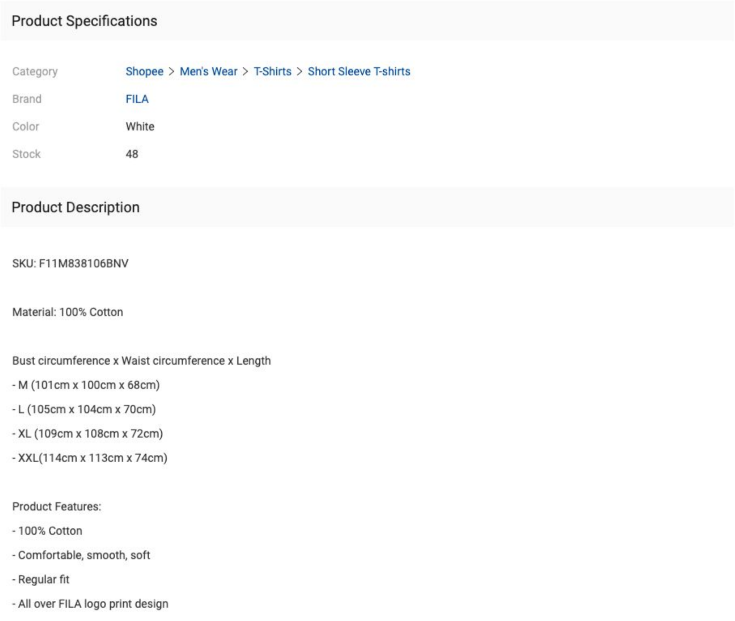
4.9 ★★★★★ | 28 Ratings | 55 Sold

~~$108.00~~ $97.00 10% OFF

Shop Vouchers	$5 OFF 10% off
Coins	Buy and earn 50 Shopee Coin
Shipping Fee	Free shipping Free shipping for orders over $20 Shipping Fee From overseas Shipping Fee $0.00
Variation	Navy M Navy L Navy XL Navy XXL
Quantity	− 1 + 14 piece available

fila_official
Active 7 Minutes Ago

Ratings	696	Response Rate	100%	Joined	15 months ago
Products	630	Response Time	within hours	Follower	40.9k



Product Specifications

Category	Shopee > Men's Wear > T-Shirts > Short Sleeve T-shirts
Brand	FILA
Color	White
Stock	48

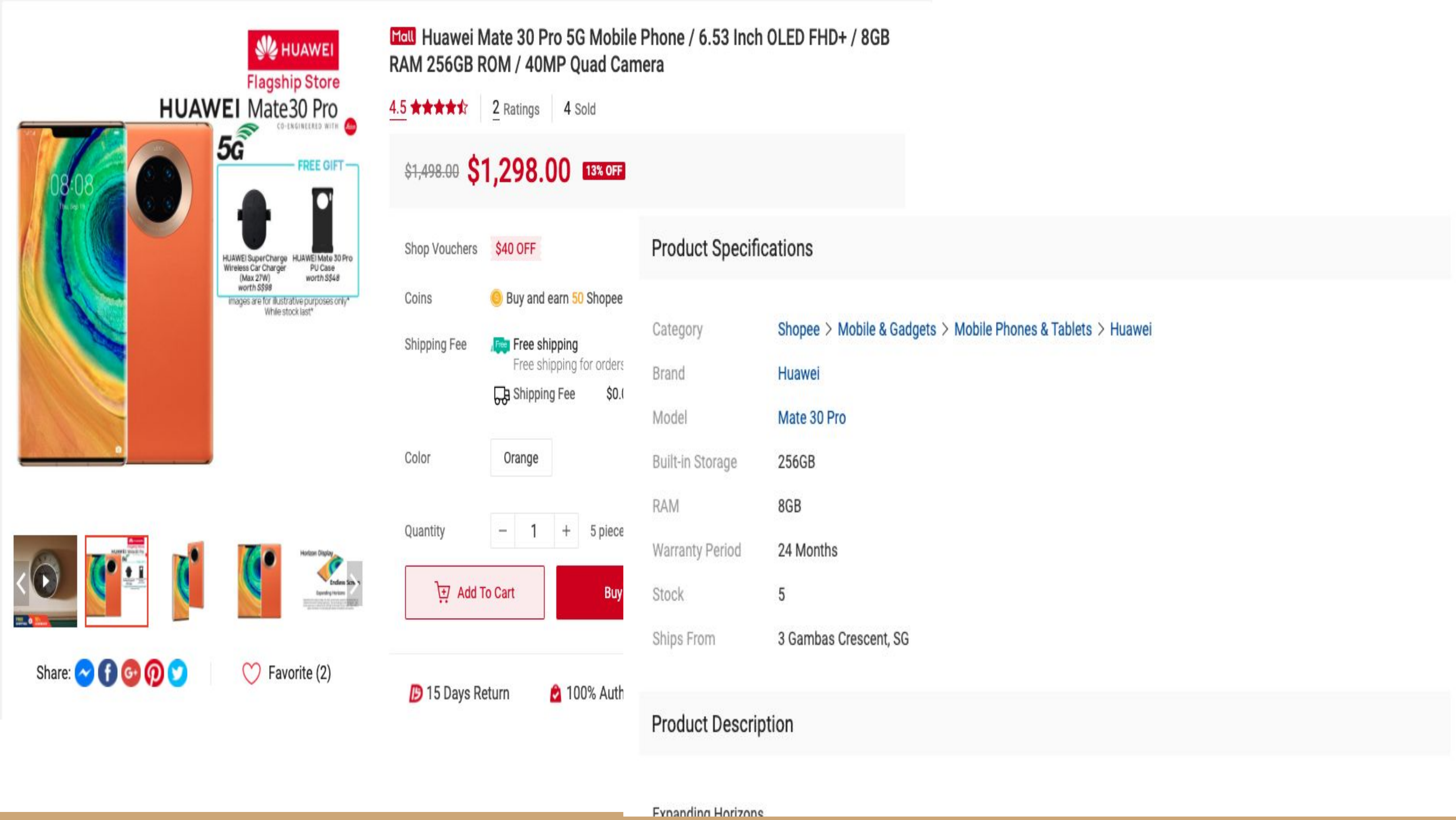
Product Description

SKU: F11M838106BNV

Material: 100% Cotton

Bust circumference x Waist circumference x Length
- M (101cm x 100cm x 68cm)
- L (105cm x 104cm x 70cm)
- XL (109cm x 108cm x 72cm)
- XXL(114cm x 113cm x 74cm)

Product Features:
- 100% Cotton
- Comfortable, smooth, soft
- Regular fit
- All over FILA logo print design


HUAWEI
Flagship Store
HUAWEI Mate30 Pro
CO-ENGINEERED WITH
5G
FREE GIFT
HUAWEI SuperCharge Wireless Car Charger (Max 27W) worth S$98
HUAWEI Mate 30 Pro PU Case worth S$48
images are for illustrative purposes only*
While stock last*
08:08
Horizon Display
Endless Scr
Share:
Favorite (2)

Mall Huawei Mate 30 Pro 5G Mobile Phone / 6.53 Inch OLED FHD+ / 8GB RAM 256GB ROM / 40MP Quad Camera

4.5 ★★★★☆ | 2 Ratings | 4 Sold

~~$1,498.00~~ $1,298.00 13% OFF

Shop Vouchers	$40 OFF
Coins	Buy and earn 50 Shopee
Shipping Fee	Free shipping Free shipping for orders Shipping Fee $0.(
Color	Orange
Quantity	− 1 + 5 piece

Add To Cart | Buy

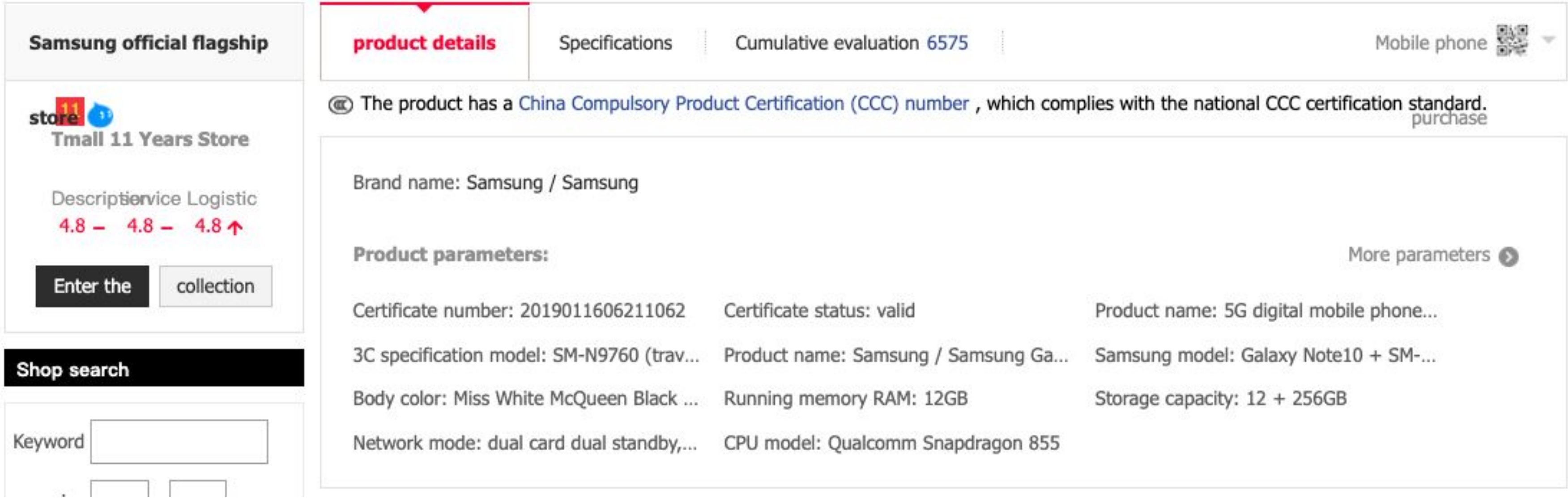
15 Days Return | 100% Auth

Product Specifications

Category	Shopee > Mobile & Gadgets > Mobile Phones & Tablets > Huawei
Brand	Huawei
Model	Mate 30 Pro
Built-in Storage	256GB
RAM	8GB
Warranty Period	24 Months
Stock	5
Ships From	3 Gambas Crescent, SG

Product Description

Expanding Horizons



Samsung official flagship

store 11

Tmall 11 Years Store

Descriptiservice Logistic

4.8 – 4.8 – 4.8 ↑

Enter the | collection

Shop search

Keyword

product details | Specifications | Cumulative evaluation 6575 | Mobile phone

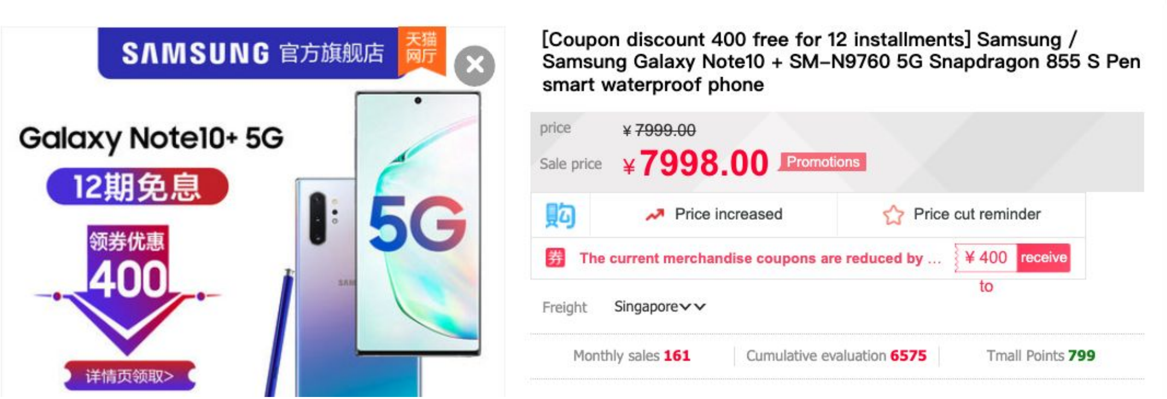
The product has a China Compulsory Product Certification (CCC) number , which complies with the national CCC certification standard.
purchase

Brand name: Samsung / Samsung

Product parameters:

More parameters

Certificate number: 2019011606211062	Certificate status: valid	Product name: 5G digital mobile phone...
3C specification model: SM-N9760 (trav...	Product name: Samsung / Samsung Ga...	Samsung model: Galaxy Note10 + SM-...
Body color: Miss White McQueen Black ...	Running memory RAM: 12GB	Storage capacity: 12 + 256GB
Network mode: dual card dual standby,...	CPU model: Qualcomm Snapdragon 855	



[Coupon discount 400 free for 12 installments] Samsung / Samsung Galaxy Note10 + SM–N9760 5G Snapdragon 855 S Pen smart waterproof phone

price ¥ ~~7999.00~~

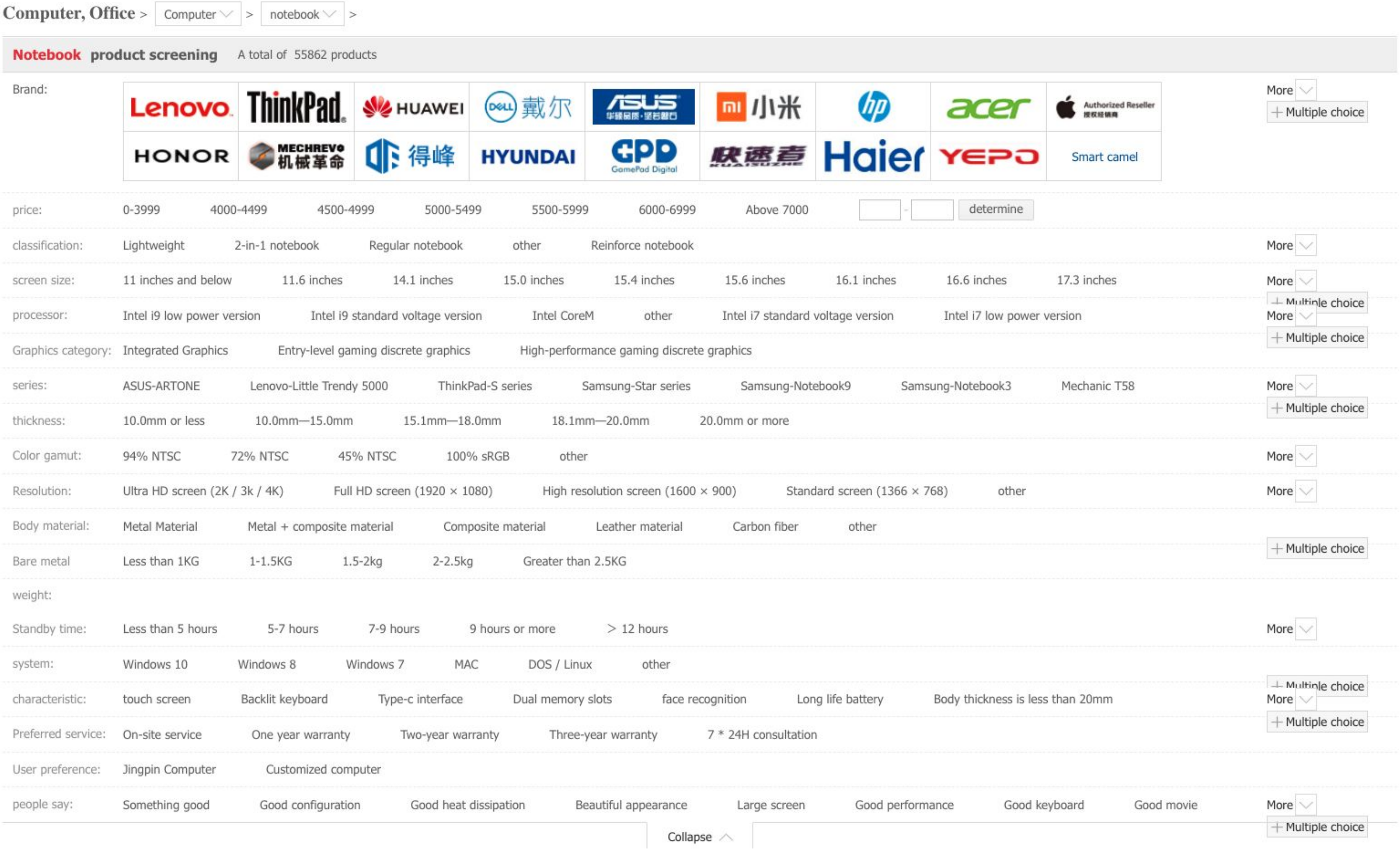
Sale price ¥7998.00 Promotions

Price increased | Price cut reminder

券 The current merchandise coupons are reduced by ... ¥ 400 receive
to

Freight Singapore

Monthly sales 161 | Cumulative evaluation 6575 | Tmall Points 799



Computer, Office > Computer > notebook >

Notebook **product screening** A total of 55862 products

Brand:	Lenovo	ThinkPad	HUAWEI	DELL 戴尔	ASUS 华硕品质·坚若磐石	mi 小米	hp	acer	Authorized Reseller 授权经销商	More + Multiple choice
	HONOR	MECHREVO 机械革命	得峰	HYUNDAI	GPD GamePad Digital	快速者 KUAISUZHE	Haier	YEPO	Smart camel	
price:	0-3999	4000-4499	4500-4999	5000-5499	5500-5999	6000-6999	Above 7000	□ - □ determine		
classification:	Lightweight	2-in-1 notebook	Regular notebook	other	Reinforce notebook					More
screen size:	11 inches and below	11.6 inches	14.1 inches	15.0 inches	15.4 inches	15.6 inches	16.1 inches	16.6 inches	17.3 inches	More + Multiple choice
processor:	Intel i9 low power version	Intel i9 standard voltage version	Intel CoreM	other	Intel i7 standard voltage version	Intel i7 low power version				More + Multiple choice
Graphics category:	Integrated Graphics	Entry-level gaming discrete graphics	High-performance gaming discrete graphics							
series:	ASUS-ARTONE	Lenovo-Little Trendy 5000	ThinkPad-S series	Samsung-Star series	Samsung-Notebook9	Samsung-Notebook3	Mechanic T58			More + Multiple choice
thickness:	10.0mm or less	10.0mm—15.0mm	15.1mm—18.0mm	18.1mm—20.0mm	20.0mm or more					
Color gamut:	94% NTSC	72% NTSC	45% NTSC	100% sRGB	other					More
Resolution:	Ultra HD screen (2K / 3k / 4K)	Full HD screen (1920 × 1080)	High resolution screen (1600 × 900)	Standard screen (1366 × 768)	other					More
Body material:	Metal Material	Metal + composite material	Composite material	Leather material	Carbon fiber	other				+ Multiple choice
Bare metal	Less than 1KG	1-1.5KG	1.5-2kg	2-2.5kg	Greater than 2.5KG					
weight:										
Standby time:	Less than 5 hours	5-7 hours	7-9 hours	9 hours or more	> 12 hours					More
system:	Windows 10	Windows 8	Windows 7	MAC	DOS / Linux	other				+ Multiple choice
characteristic:	touch screen	Backlit keyboard	Type-c interface	Dual memory slots	face recognition	Long life battery	Body thickness is less than 20mm			More + Multiple choice
Preferred service:	On-site service	One year warranty	Two-year warranty	Three-year warranty	7 * 24H consultation					
User preference:	Jingpin Computer	Customized computer								
people say:	Something good	Good configuration	Good heat dissipation	Beautiful appearance	Large screen	Good performance	Good keyboard	Good movie		More + Multiple choice

Collapse

Why attributes extraction is fundamental in e-commercial ?

- Auto fill in attributes specification

- Improve search and recommendation

- Build product graph

2.2 Data collection

- Define the data scope
- Define the attribute types
- Define initial values in each attribute type

Material	Pattern	Neckline	Sleeve Length	Top Fit Type	Pant Fit Type	Pant Length
Denim, Cotton, Leather, Polyester, Other Material, Spandex, Wool	Checkered(Plaid), Print, Floral, Other Pattern, Plain, Striped, Tie Dye, Polka Dotted	V Neck, Round Neck, Turtle Neck, Henley, Other Neckline	Short Sleeves, ¾ Sleeves, Long Sleeves, Sleeveless	Slim Fit, Regular Fit, Relax Fit,	Slim Fit, Regular Fit, Relax Fit	Ankle, Full Length

We may use above attribute types and values in men fashion

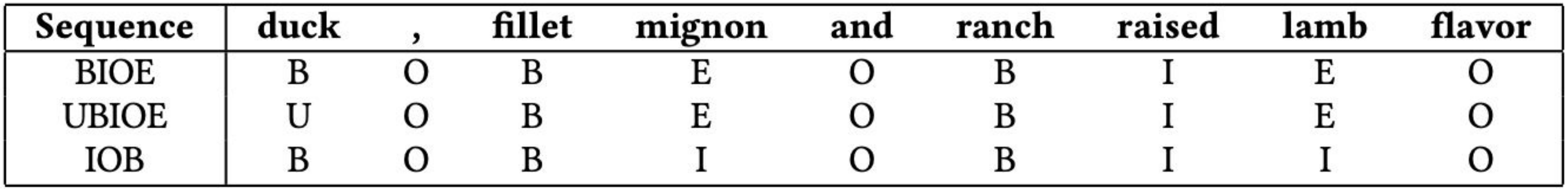
2.2 Data collection

Collect data from Shopee backend database including:

- Product title
- Product seller input attributes
- Product description

All data is free text (unstructured data)

2.2 Data collection



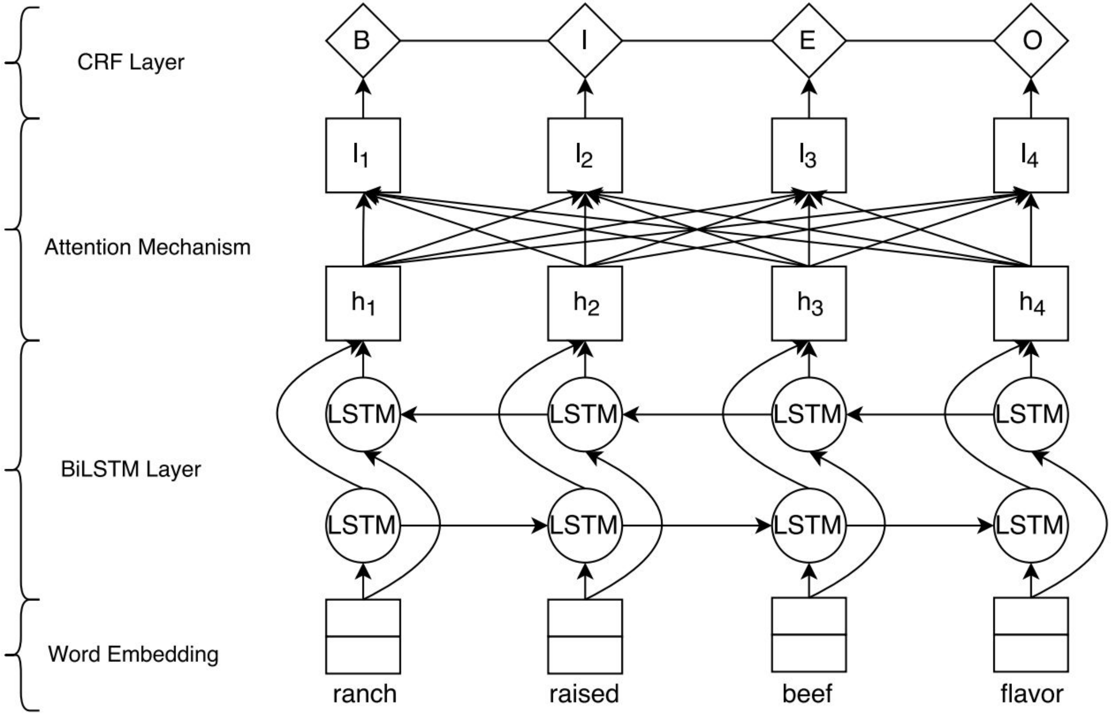
Sequence	duck	,	fillet	mignon	and	ranch	raised	lamb	flavor
BIOE	B	O	B	E	O	B	I	E	O
UBIOE	U	O	B	E	O	B	I	E	O
IOB	B	O	B	I	O	B	I	I	O

In **BIOE** tagging strategy,

- **B** :represents the beginning of an attribute,
- **I** :represents the inside of an attribute,
- **O** :represents the outside of an attribute,
- **E** :represents the end of an attribute.

“**UBIOE**" has an extra tag ‘**U**’ representing the unit token tag that separates one-word attributes from multi-word ones.

2.3 Baseline Model



Attention Mechanism:

$$g_{t,t'} = tanh(W_g h_t + W_{g'} h_{t'} + b_g),$$

$$\alpha_{t,t'} = \sigma(W_a g_{t,t'} + b_a),$$

2.4 Pain Points

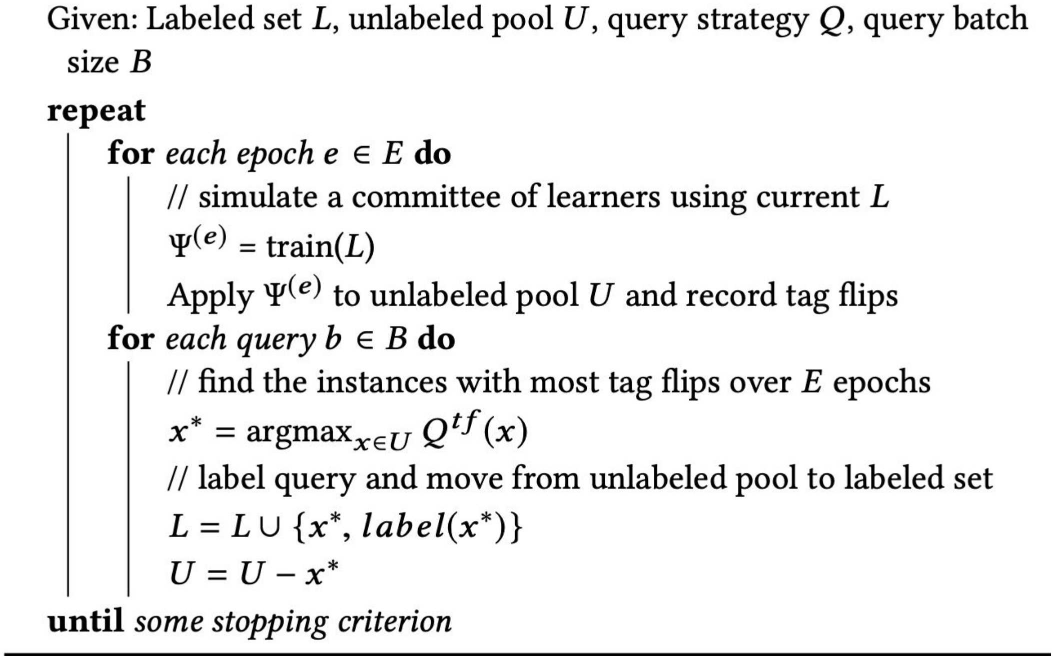
- Only limited data is well labelled
  - Matching data with Vocabulary is cheap but not good
  - Manual annotation costs much time, hard to annotate and keep high quality
  - Cold start to expand training data
- Hard to evaluate the model performance
  - Limited ground true test data
  - can we develop models that give interpretable explanation for its decisions, unlike black-box methods that are difficult to debug?
  - Only F1 score (precision and recall) can not meet business requirements
- Hard to improve the model generality ability
  - Performs pool in extracting unseen attribute values
  - Too close to exact matching with vocabularies

2.5 Conclusions and future work

- NER model's recall is important
    a. Expand existing attributes dictionary
    b. Open World Assumption (OWA) is common in e-commercial website
- Low resource NER:
    - Active learning
    - Multi task training
    - Transfer learning

Active learning in NER

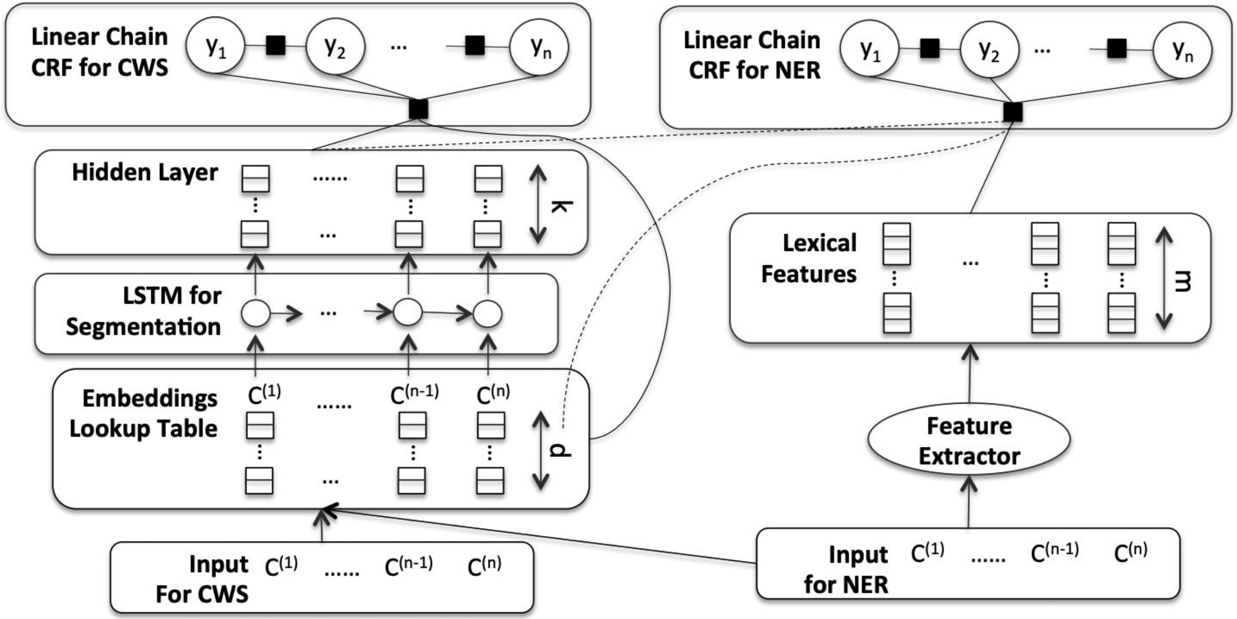
example: Active learning with tag flips as query strategy



```
Given: Labeled set $L$, unlabeled pool $U$, query strategy $Q$, query batch
  size $B$
repeat
    for each epoch $e \in E$ do
        // simulate a committee of learners using current $L$
        $\Psi^{(e)}$ = train($L$)
        Apply $\Psi^{(e)}$ to unlabeled pool $U$ and record tag flips
    for each query $b \in B$ do
        // find the instances with most tag flips over $E$ epochs
        $x^* = \text{argmax}_{x \in U}\, Q^{tf}(x)$
        // label query and move from unlabeled pool to labeled set
        $L = L \cup \{x^*, label(x^*)\}$
        $U = U - x^*$
until some stopping criterion
```

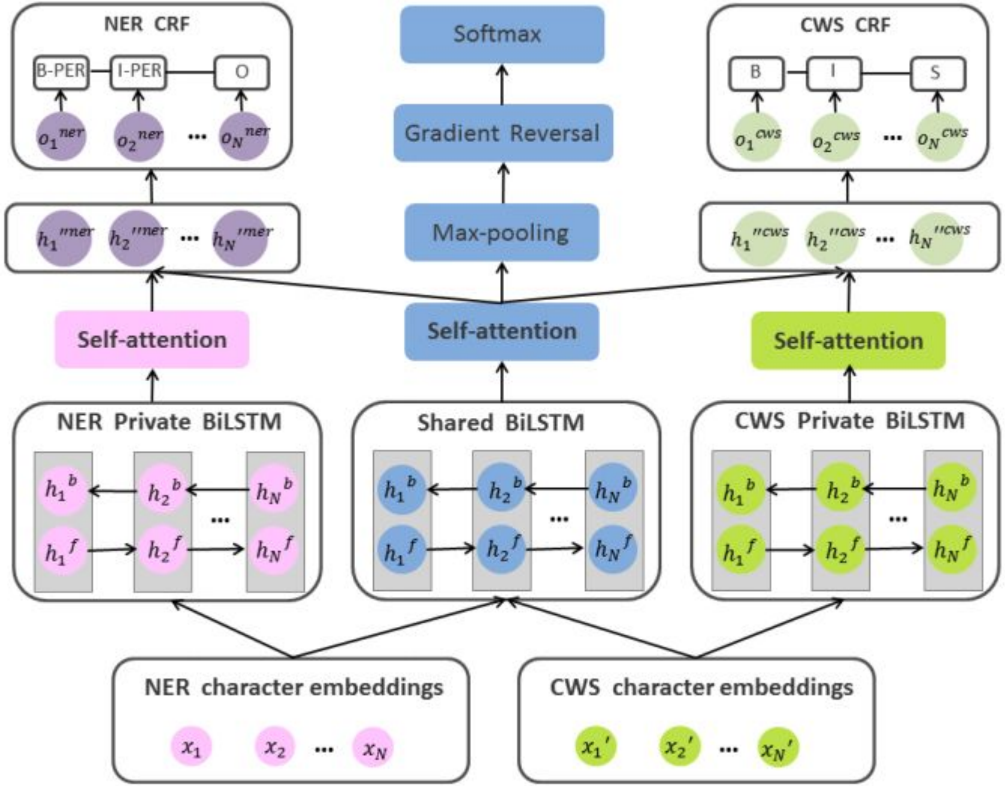
Multi task learning in NER

example: Jointly training NER and word segmentation with an LSTM-CRF model



For languages where word boundaries are not readily identified in text, word segmentation is a key first step to generating features for an NER system. While using word boundary tags as features are helpful, the signals that aid in identifying these boundaries may provide richer information for an NER system

Transfer learning in NER


NER CRF
B-PER
I-PER
O
o_1^{ner} o_2^{ner} ... o_N^{ner}
$h_1''^{ner}$ $h_2''^{ner}$ ... $h_N''^{ner}$
Self-attention
NER Private BiLSTM
h_1^b h_2^b h_N^b
h_1^f h_2^f h_N^f
Softmax
Gradient Reversal
Max-pooling
Self-attention
Shared BiLSTM
CWS CRF
B
I
S
o_1^{cws} o_2^{cws} ... o_N^{cws}
$h_1''^{cws}$ $h_2''^{cws}$ ... $h_N''^{cws}$
Self-attention
CWS Private BiLSTM
NER character embeddings
x_1 x_2 ... x_N
CWS character embeddings
x_1' x_2' ... x_N'