



CISC3025 Natural Language Processing

# Part-Of-Speech Tagging with Viterbi Algorithm

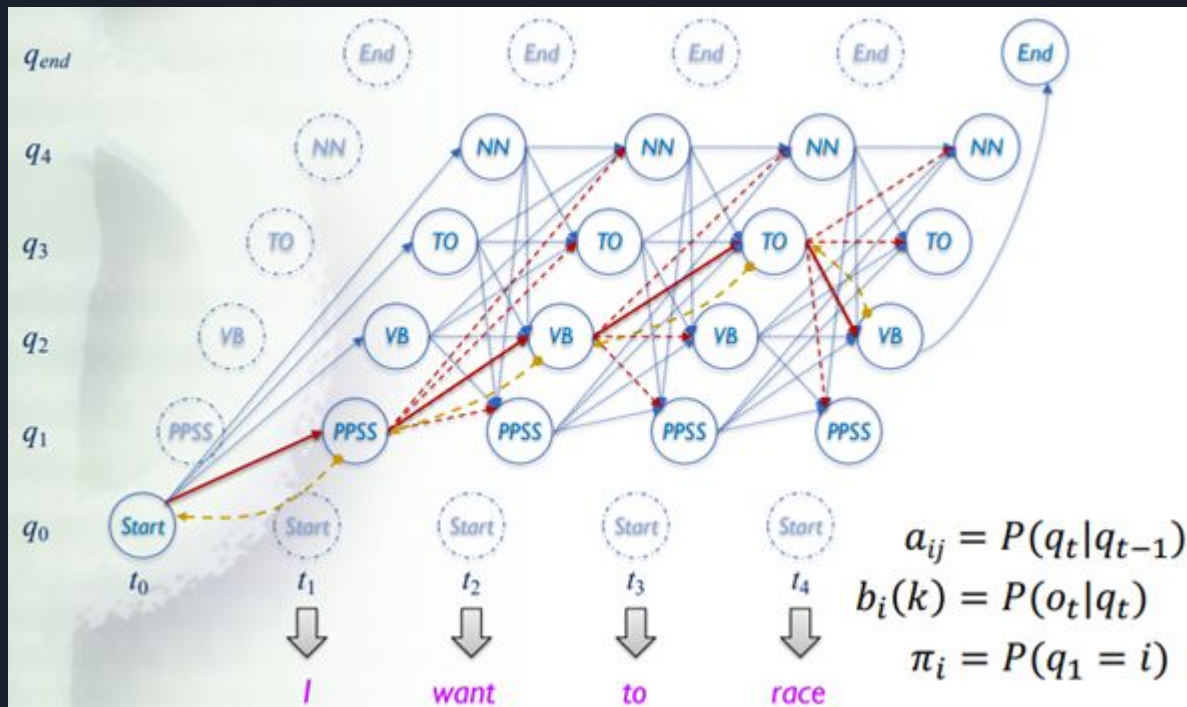
DC027649 Jiang Rui  
DC027857 Wang XinYu



# Table of Contents

1. Viterbi algorithm
2. Feature selection
3. Chi-Square
4. Model training and testing
5. Model evaluation
6. Conclusion

# 1. Viterbi algorithm



`viterbi[label][word]=max (P(label | token) * P(prev_label | label)) * viterbi[prev_label][prev_token]`

## 2. Features selection

1. Position in the sentence
  - a. is the first word / is the last word
2. Components
  - a. contains numbers/punctuations/capital letters
  - b. consists of numbers/punctuations/capital letters
3. Morphology
  - a. has prefix of un-/in-/pre-/dis-...
  - b. has suffix of -ed/-ly/-ing/-tion...

$$\text{viterbi}[\text{label}][\text{word}] = \max (\text{P}(\text{label} \mid \text{token}) * \text{P}(\text{prev\_label} \mid \text{label})) * \\ \text{viterbi}[\text{prev\_label}][\text{prev\_token}] * \Pi \text{P}(\text{feature} \mid \text{label})$$


### 3. The Chi-square test

The Chi-square test is often used to analyze the relevance between two classification variables

Feature\Tag	Tag A	Tag B
Feature X	3	3
Feature Y	6	0

Chi-square value: Feature Y > Feature X

## 4. Model training and testing — word vectorization



Word2Vector is to convert a word into a feature vector, which is a sequence of feature values of the word.

Example:


sentence: I love you.

features: 'is\_punctuation', 'is\_first\_word', 'is\_last\_word' and 'is\_complete\_capital'

Word "I": is\_punctuation = 0, is\_first\_word = 1, is\_last\_word = 0, is\_complete\_capital = 1

**"I" → (0, 1, 0, 1)**

## 4. Model training and testing — word vectorization



	<i>is_punctuation</i>	is_first_word	is_last_word	is_complete_capital
I	0	1	0	1
love	0	0	0	0
you	0	0	0	0
.	1	0	1	0

“I love you.” → (0, 1, 0, 1), (0, 0, 0, 0), (0, 0, 0, 0), (1, 0, 1, 0)

(“I love you. ”), (“I hate you.”)

→ (0, 1, 0, 1), (0, 0, 0, 0), (0, 0, 0, 0), (1, 0, 1, 0), (0, 1, 0, 1), (0, 0, 0, 0), (0, 0, 0, 0), (1, 0, 1, 0)

## 4. Model training and testing — train model matrices

1. The probability of each word being labeled as each label:  $P(\text{word} \mid \text{label})$
2. The probability of previous labels per pair:  $P(\text{prev\_label} \mid \text{label})$
3. The probability of each feature appearing in each label:  $P(\text{feature} \mid \text{label})$
4. The probability of each label appearing in the training set:  $P(\text{label})$

Example: matrix of feature suffix\_-tion on label (NN, IN, DT)

	NN	IN	DT
0	0.0682	1	1
1	0.9318	0	0




## 4. Model training and testing — prediction

For a input sentence, construct a Viterbi matrix for each word to predict  $P(\text{label} \mid \text{word})$

$$\text{viterbi}[\text{label}][\text{word}] = \max P(\text{prev\_label} \mid \text{label})) * \text{viterbi}[\text{prev\_label}][\text{prev\_token}] \\ * (P(\text{label} \mid \text{token}) * \prod P(\text{feature} \mid \text{label}))$$

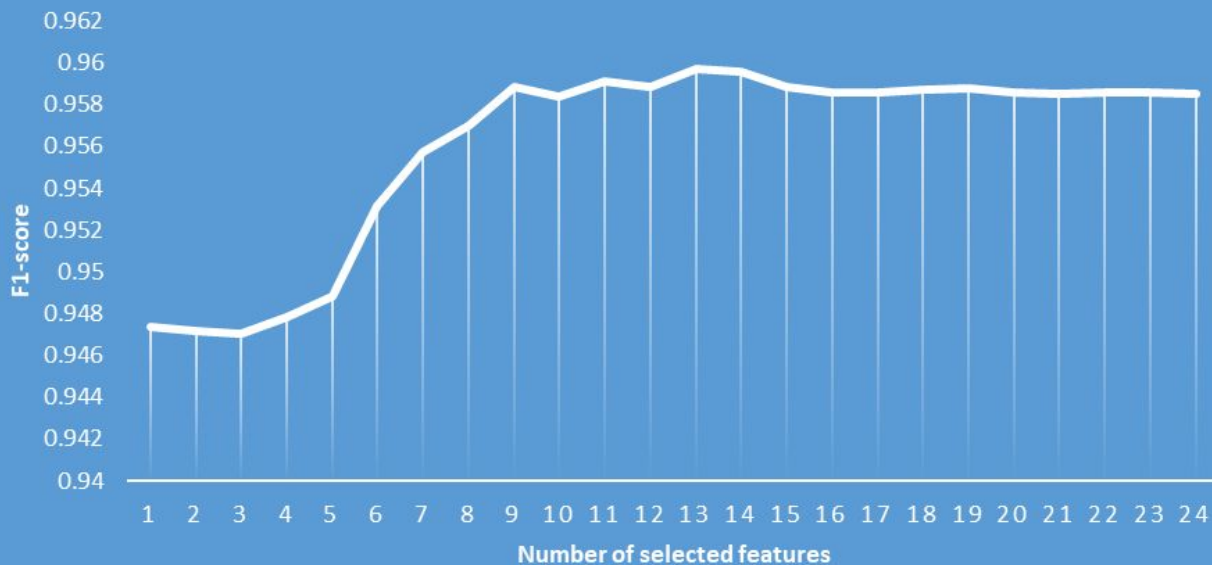
	I	got	smoke
NN	0.02	0.000003	0.0000004
IN	0.0004	0.000002	0.00000002
DT	0.0002	0.000004	0.00000008



I – NN, got – IN, smoke – NN

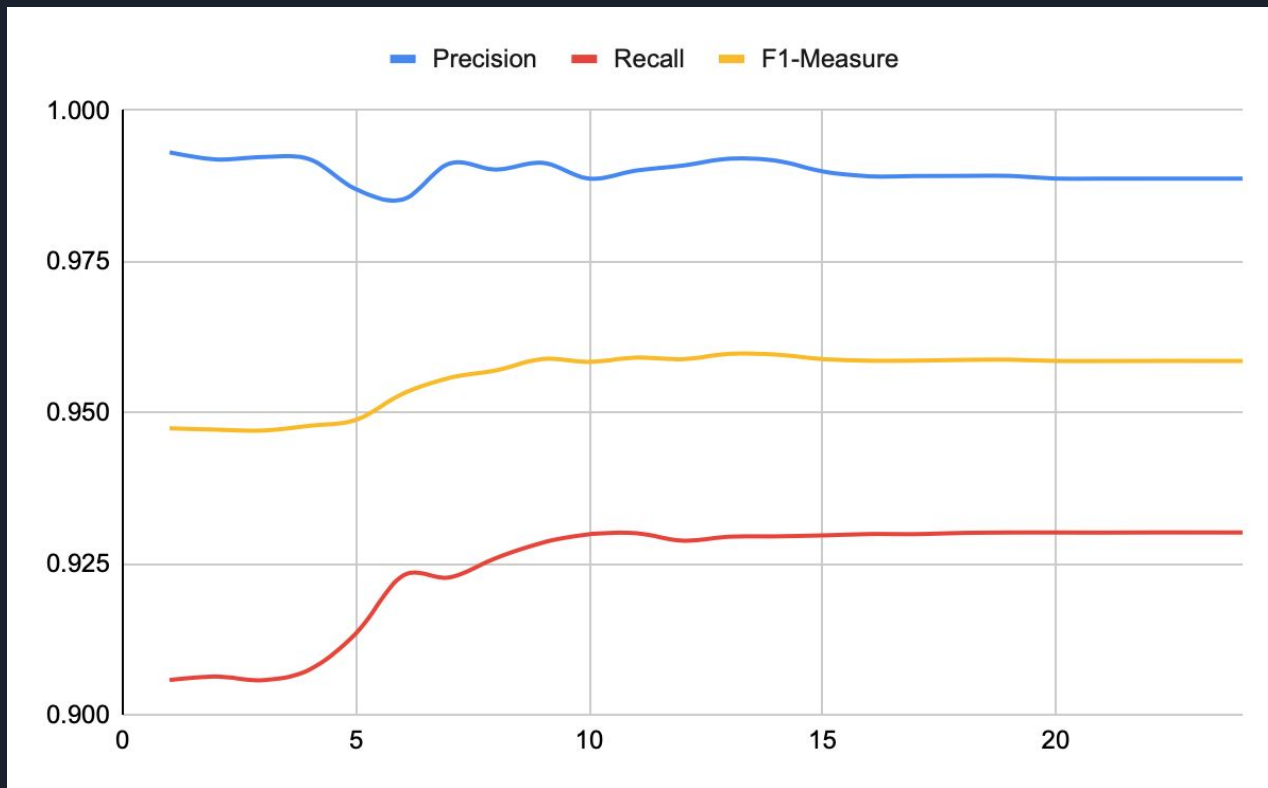
## 5. Model evaluation

Micro F1-measure is used to evaluate the performance of model



## 5. Model evaluation

Micro F1-measure is used to evaluate the performance of model




## 5. Model evaluation

Most frequency POS labels in test set and their confusion matrix

POS	TP	FN	FP	Count	Precision	Recall	F1-measure
NN	14951	1206	46	16203	0.92535743	0.99693272	0.95981254
IN	13016	227	49	13292	0.98285887	0.99624952	0.98950889
DT	11465	118	47	11630	0.98981266	0.9959173	0.9928556
JJ	6621	1279	48	7948	0.83810127	0.99280252	0.90891619
NNP	6401	190	48	6639	0.97117281	0.99255699	0.98174847

## 6. Conclusion and guess

- 
1. A larger training set would improve the performance of model greatly.
  2. It is not the case that with more features added into the model, the predict accuracy would be higher. Only first 5- 10 features added would increase accuracy significantly. Maybe it is due to the issue of overfit if too many features included.
  3. The selection of feature would directly influence prediction accuracy of certain labels, if we want to further increase prediction accuracy, maybe more specific effective features for low accuracy labels are needed.



# Ends

Thanks for listening