



# 文字探勘 (Text Mining)

## 文字探勘課程介紹 (Course Orientation on Text Mining)

1082TM01

MBA, BDABI, TKU (E3611) (8480) (Spring 2020)

Mon, 7, 8, 9 (14:10-17:00) (B206)



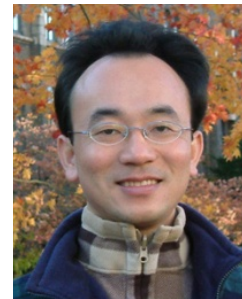
Chichang Jou

周清江

Associate Professor

副教授

[cjou@mail.tku.edu.tw](mailto:cjou@mail.tku.edu.tw)



Min-Yuh Day

戴敏育

Associate Professor

副教授

[myday@mail.tku.edu.tw](mailto:myday@mail.tku.edu.tw)

Dept. of Information Management, Tamkang University

淡江大學 資訊管理學系

# 文字探勘 (Text Mining)

# 淡江大學108學年度第2學期

## 課程教學計畫表

Spring 2020 (2020.02 - 2020.06)

- 課程名稱：**文字探勘 (Text Mining)**
- 授課教師：周清江 (Chichang Jou) ，  
戴敏育 (Min-Yuh Day)
- 開課系級：大數據碩士學程 (TLXDM)  
Master's Program in Big Data Analytics and Business Intelligence
- 開課資料：選修 單學期 3 學分 (3 Credits, Elective)
- 上課時間：週一 7, 8, 9 (Mon 14:10-17:00)
- 上課教室：B206 (淡江大學淡水校園)

# 淡江大學大數據所 系(所)教育目標



- 培育學生具**研究大數據**的能力。
- 培育學生具**大數據程式設計**的能力。

# 淡江大學大數據所 系(所)核心能力



- A. 具**研究大數據分析理論**的能力。  
(比重：40.00)
- B. 具**大數據分析**的能力。  
(比重：40.00)
- C. 具**整合各領域之知識**的能力。  
(比重：20.00)

# 課程簡介

- 本課程介紹文字探勘基本概念與研究議題。
- 課程內容包括
  - 文字探勘的基礎：自然語言處理 (NLP)、
  - Python自然語言處理、
  - 處理和理解文本、
  - 文本表達特徵工程、
  - 文本分類、
  - 文本摘要和主題模型、
  - 文本相似度和分群、
  - 語意分析與命名實體識別 (NER)、
  - 情感分析、
  - 深度學習和通用句子嵌入模型、
  - 問答系統與對話系統、
  - 和文字探勘個案研究。

# Course Introduction

- This course introduces the fundamental concepts and research issues of Text Mining.
- Topics include
  - Foundations of Text Mining: Natural Language Processing (NLP),
  - Python for NLP,
  - Processing and Understanding Text,
  - Feature Engineering for Text Representation,
  - Text Classification,
  - Text Summarization and Topic Models,
  - Text Similarity and Clustering,
  - Semantic Analysis and Named Entity Recognition,
  - Sentiment Analysis,
  - The Promise of Deep Learning and Universal Sentence-Embedding Models,
  - Question Answering and Dialogue Systems,
  - and Case Study on Text Mining.

# 課程目標 (Objective)

- 瞭解及應用文字探勘基本概念與研究議題。  
Understand and apply the fundamental concepts and research issues of Text Mining.
- 進行文字探勘相關之資訊管理研究。  
Conduct information systems research in the context of Text Mining.



# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2020/03/02	文字探勘課程介紹 (Course Orientation on Text Mining)
2	2020/03/09	文字探勘基礎：自然語言處理 (Foundations of Text Mining: Natural Language Processing; NLP)
3	2020/03/16	Python自然語言處理 (Python for Natural Language Processing)
4	2020/03/23	處理和理解文本 (Processing and Understanding Text)
5	2020/03/30	文本表達特徵工程 (Feature Engineering for Text Representation)
6	2020/04/06	人工智慧文本分析個案研究 I (Case Study on Artificial Intelligence for Text Analytics I)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2020/04/13	文本分類 (Text Classification)
8	2020/04/20	文本摘要和主題模型 (Text Summarization and Topic Models)
9	2020/04/27	期中報告 (Midterm Project Report)
10	2020/05/04	文本相似度和分群 (Text Similarity and Clustering)
11	2020/05/11	語意分析和命名實體識別 (Semantic Analysis and Named Entity Recognition; NER)
12	2020/05/18	情感分析 (Sentiment Analysis)

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2020/05/25	人工智慧文本分析個案研究 II (Case Study on Artificial Intelligence for Text Analytics II)
14	2020/06/01	深度學習和通用句子嵌入模型 (Deep Learning and Universal Sentence-Embedding Models)
15	2020/06/08	問答系統與對話系統 (Question Answering and Dialogue Systems)
16	2020/06/15	期末報告 I (Final Project Presentation I)
17	2020/06/22	期末報告 II (Final Project Presentation II)
18	2020/06/29	教師彈性補充教學

# 教學方法與評量方法

- 教學方法

- 講述、討論、  
發表、實作

- 評量方法

- 討論、實作、報告

# 教材課本

- 教材課本
  - 講義 (Slides)
  - 文字探勘相關個案與論文  
(Cases and Papers related to Text Mining)

# 參考書籍 (References)

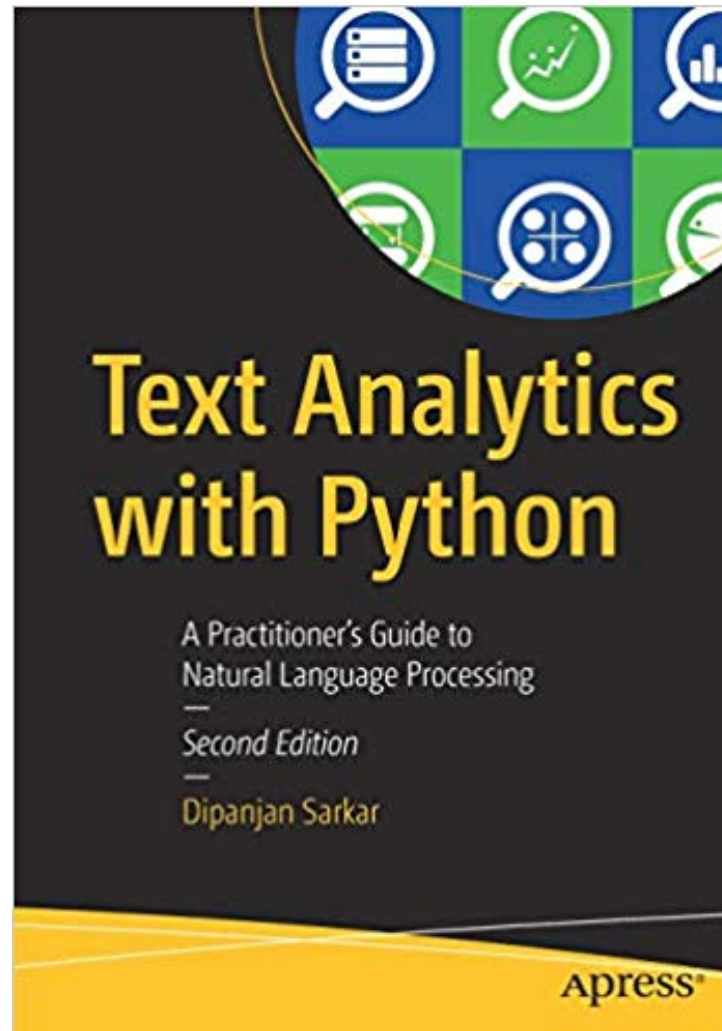
1. Dipanjan Sarkar (2019),  
Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.
2. Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),  
Applied Text Analysis with Python:  
Enabling Language-Aware Data Products with Machine Learning, O'Reilly.
3. Charu C. Aggarwal (2018),  
Machine Learning for Text, Springer.
4. Gabe Ignatow and Rada F. Mihalcea (2017),  
An Introduction to Text Mining: Research Design, Data Collection, and Analysis, SAGE Publications.

# 作業與學期成績計算方式

- 作業篇數
  - 3篇
- 學期成績計算方式
  - 期中評量：30 %
  - 期末評量：30 %
  - 其他（課堂參與及報告討論表現）：40 %

Dipanjan Sarkar (2019),

**Text Analytics with Python:**  
**A Practitioner's Guide to Natural Language Processing,**  
Second Edition. APress.

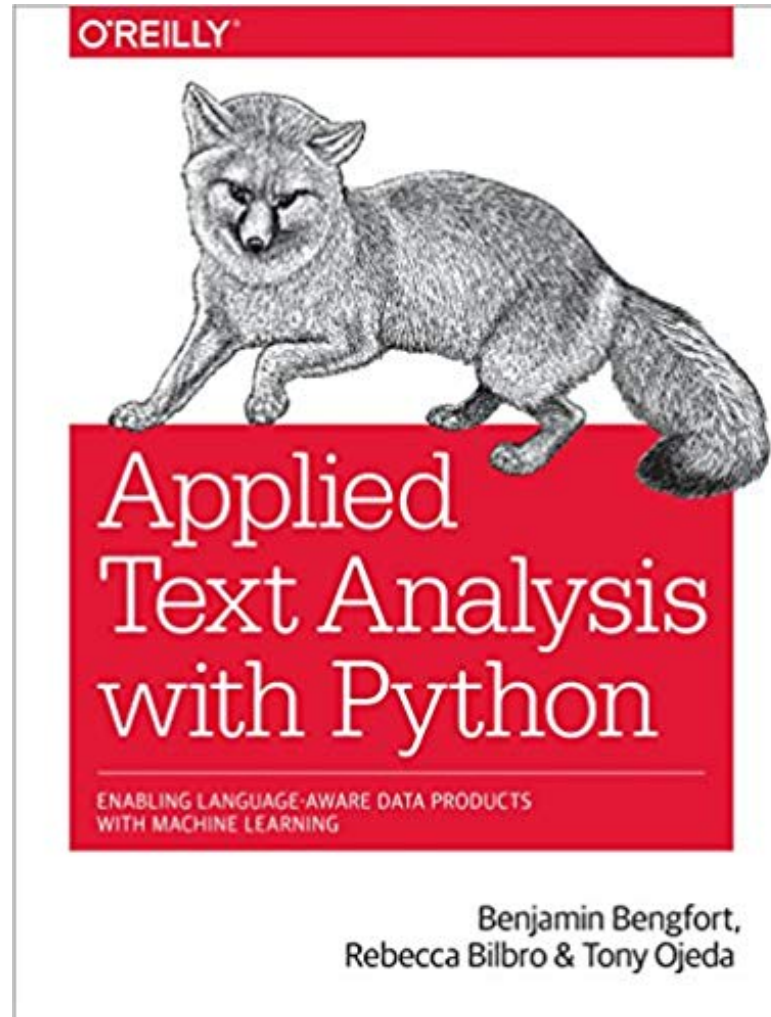




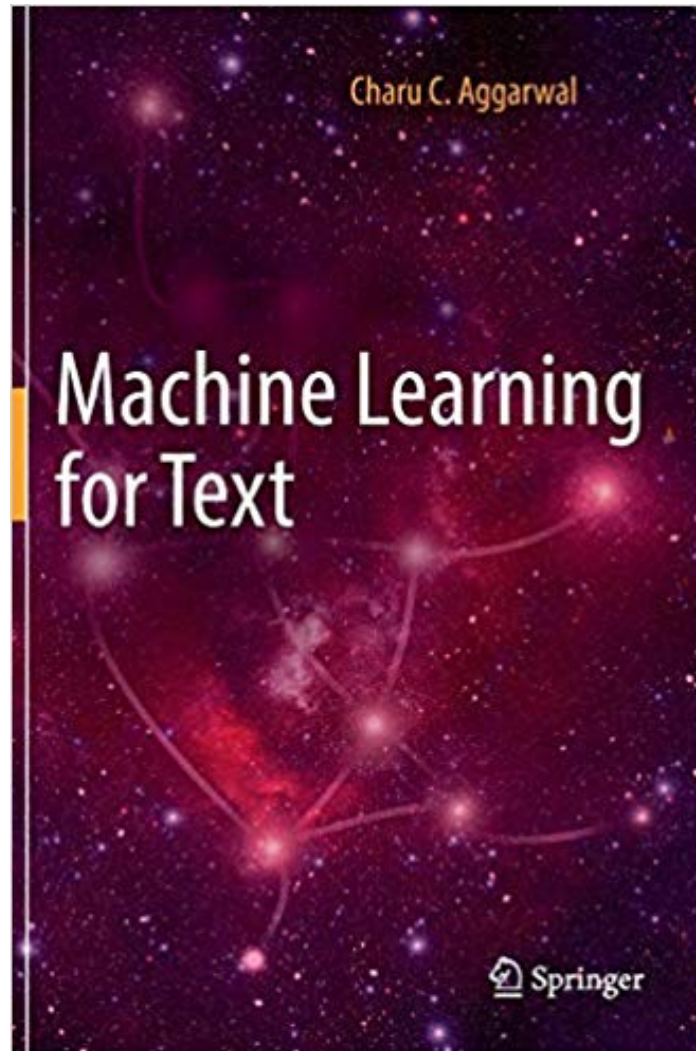
Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),

# Applied Text Analysis with Python:

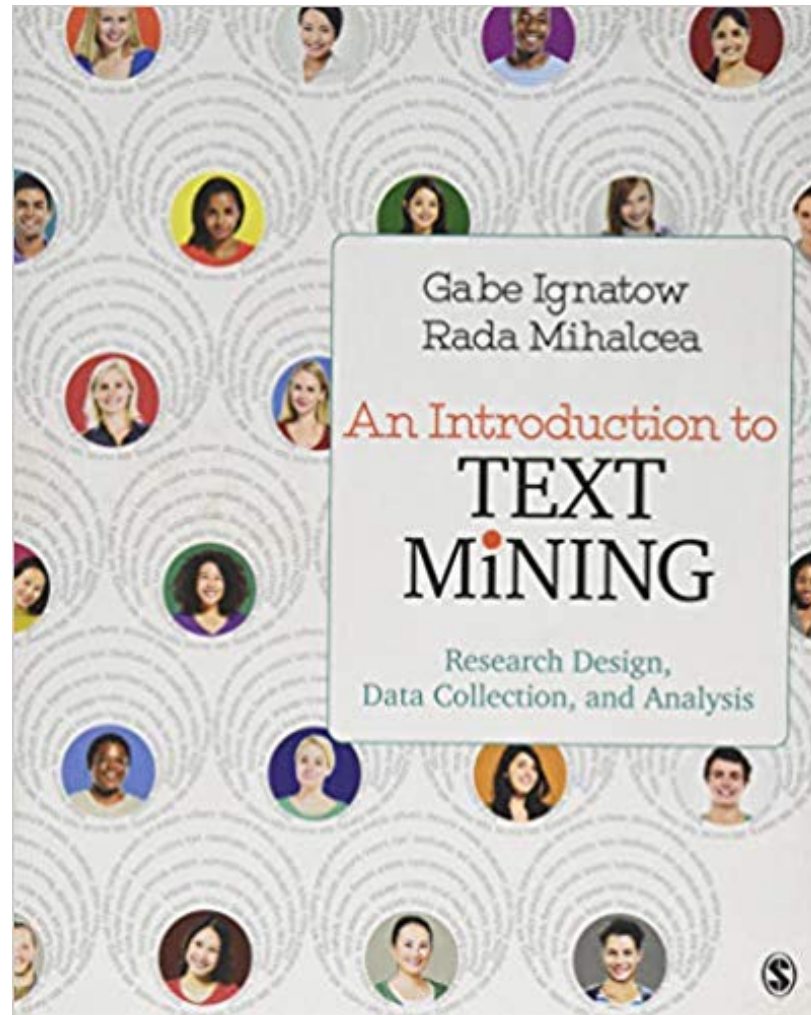
Enabling Language-Aware Data Products with Machine Learning,  
O'Reilly.



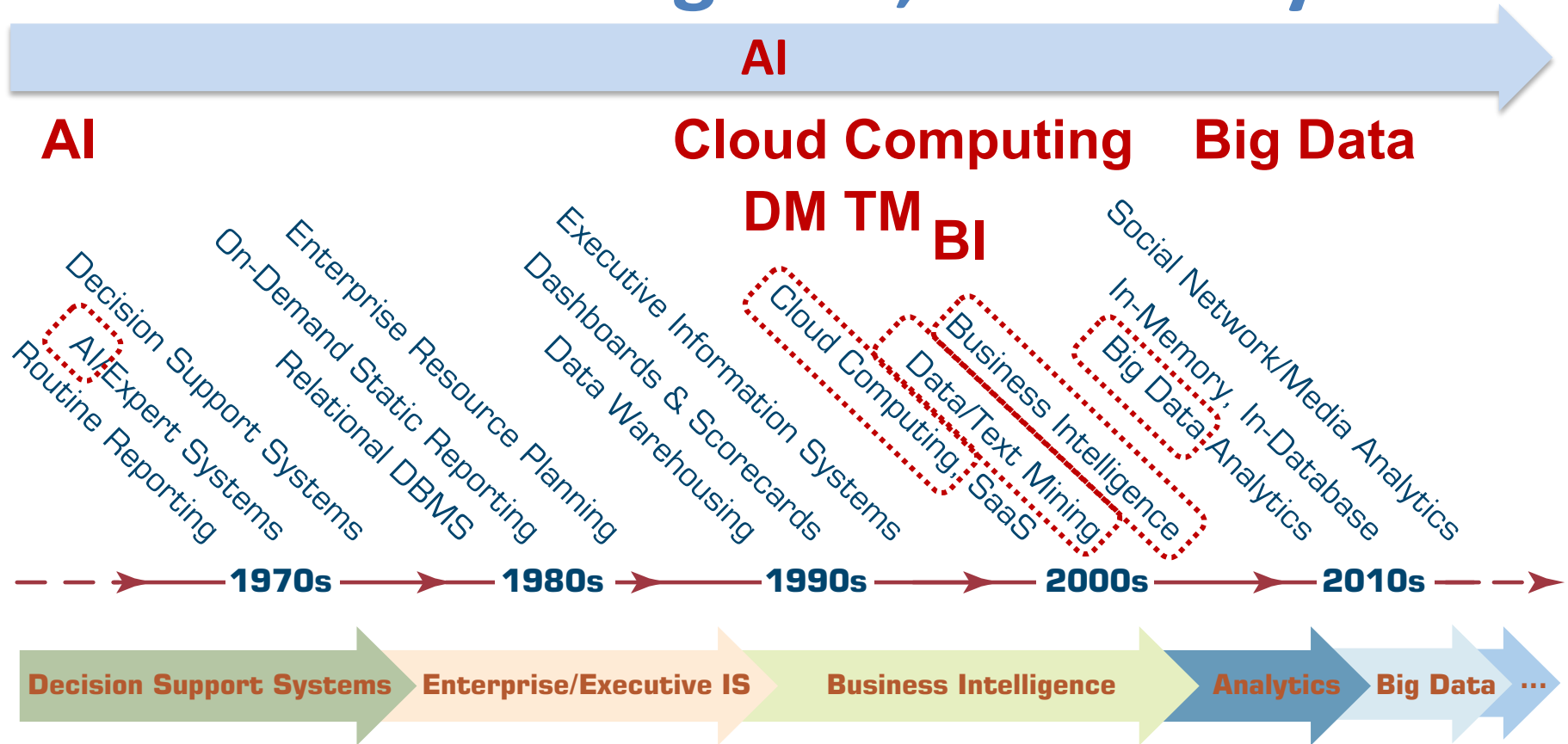
Charu C. Aggarwal (2018),  
**Machine Learning for Text**,  
Springer



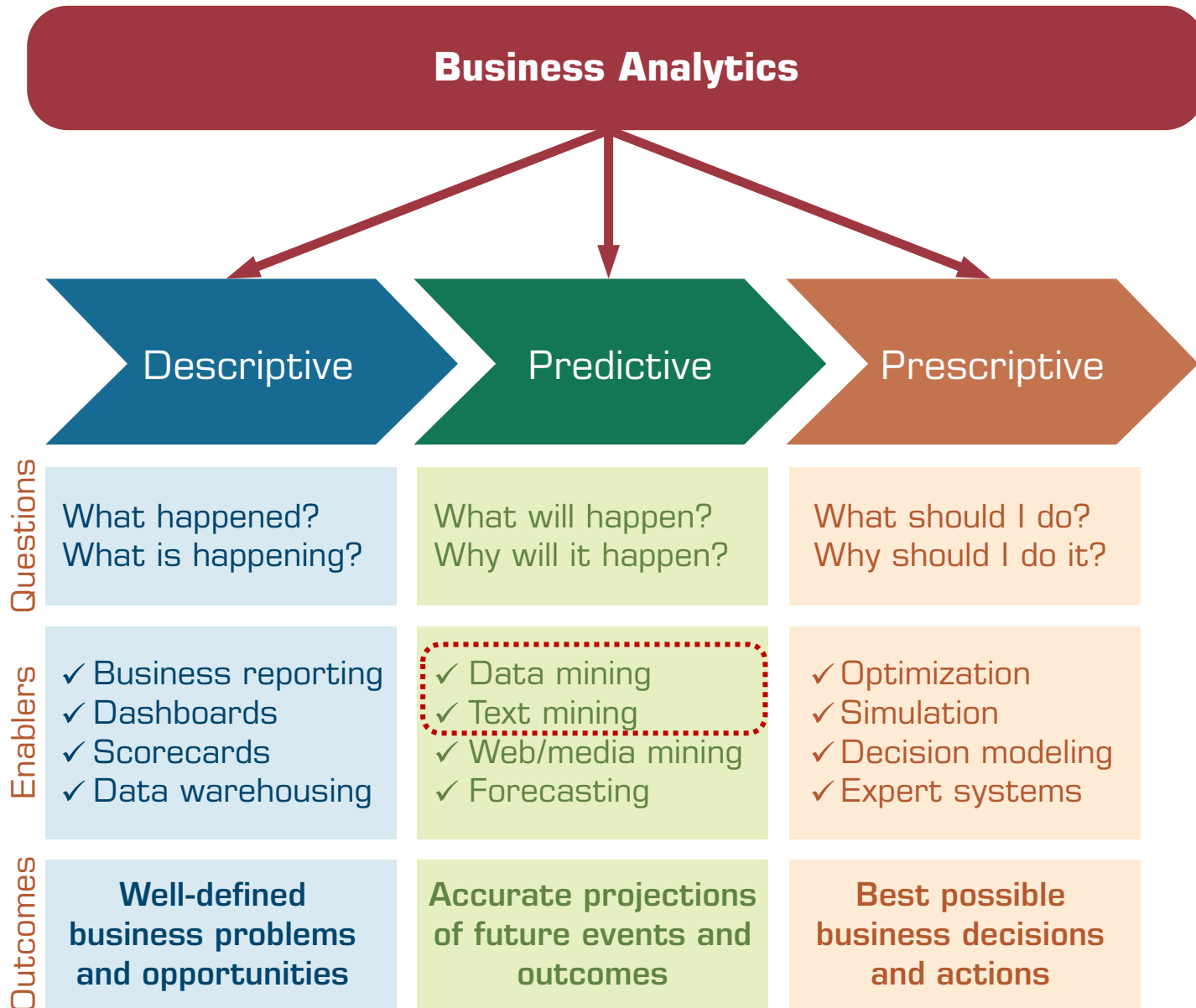
Gabe Ignatow and Rada F. Mihalcea (2017),  
**An Introduction to Text Mining:**  
**Research Design, Data Collection, and Analysis,**  
SAGE Publications.



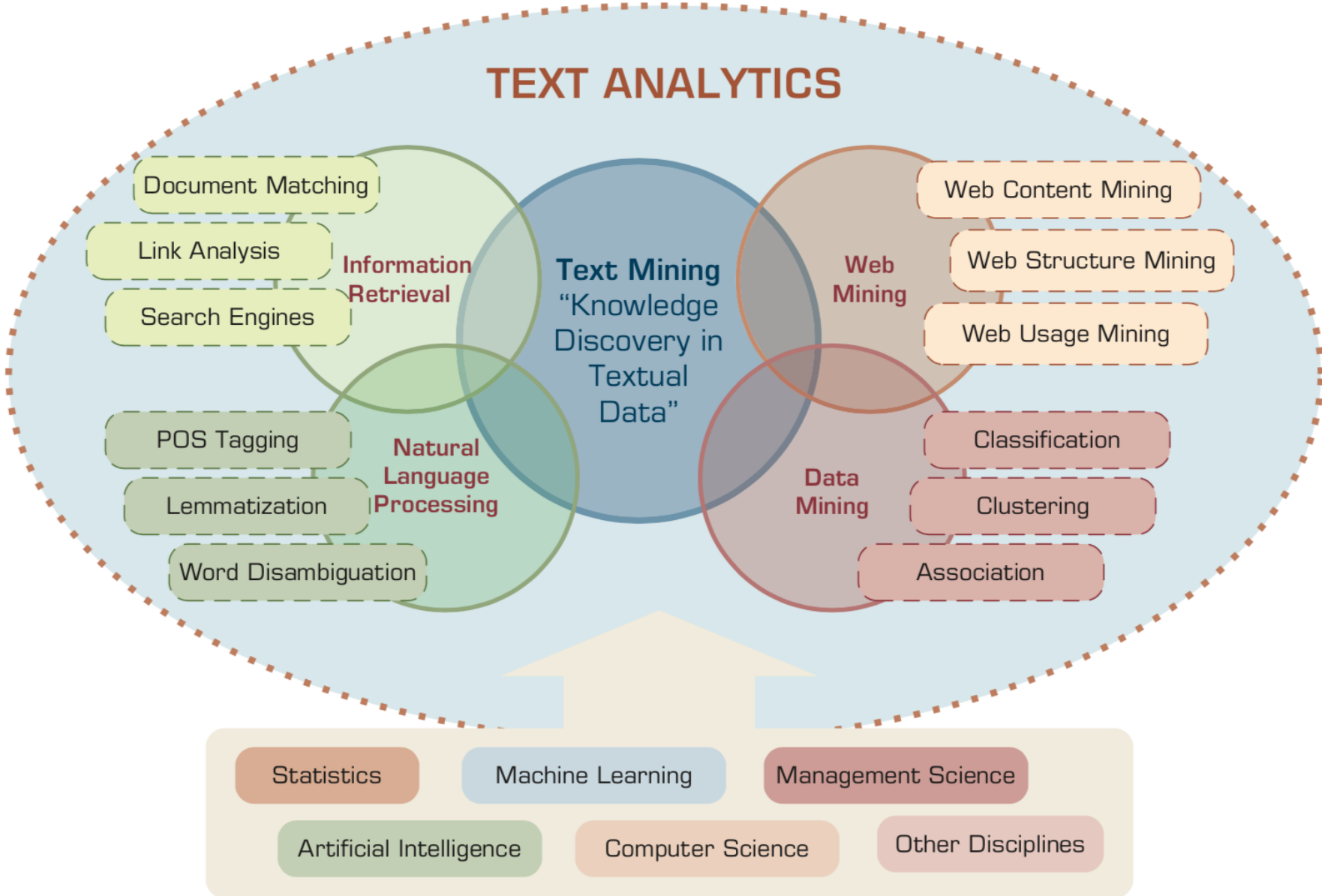
# AI, Big Data, Cloud Computing Evolution of Decision Support, Business Intelligence, and Analytics



# Three Types of Analytics



# Text Analytics and Text Mining



# Text Analytics

- **Text Analytics** =  
Information Retrieval +  
Information Extraction +  
Data Mining +  
Web Mining
- **Text Analytics** =  
Information Retrieval +  
Text Mining

# Text mining

- Text Data Mining
- Knowledge Discovery in Textual Databases



# Application Areas of Text Mining

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

# Natural Language Processing (NLP)

- Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics.

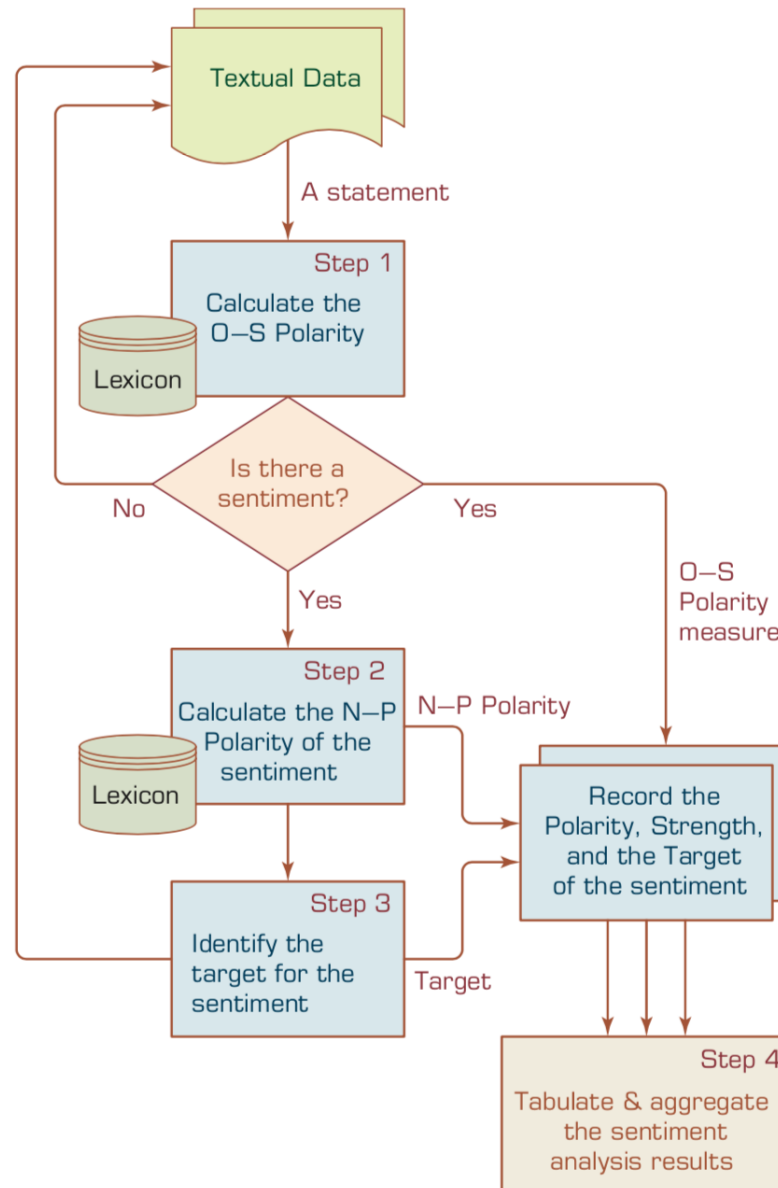
# Natural Language Processing (NLP)

- Part-of-speech tagging
- Text segmentation
- Word sense disambiguation
- Syntactic ambiguity
- Imperfect or irregular input
- Speech acts

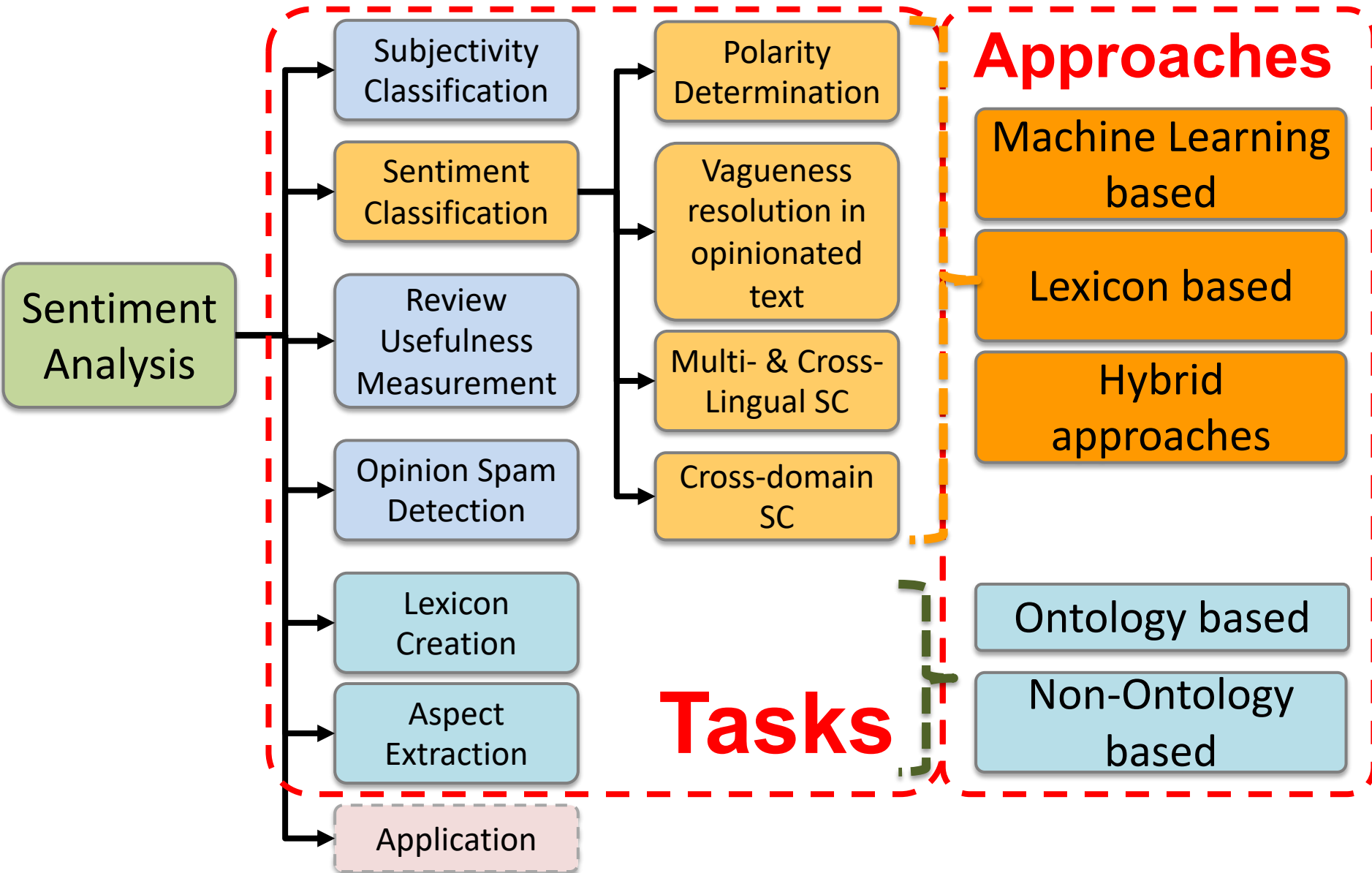
# NLP Tasks

- Question answering
- Automatic summarization
- Natural language generation
- Natural language understanding
- Machine translation
- Foreign language reading
- Foreign language writing.
- Speech recognition
- Text-to-speech
- Text proofing
- Optical character recognition

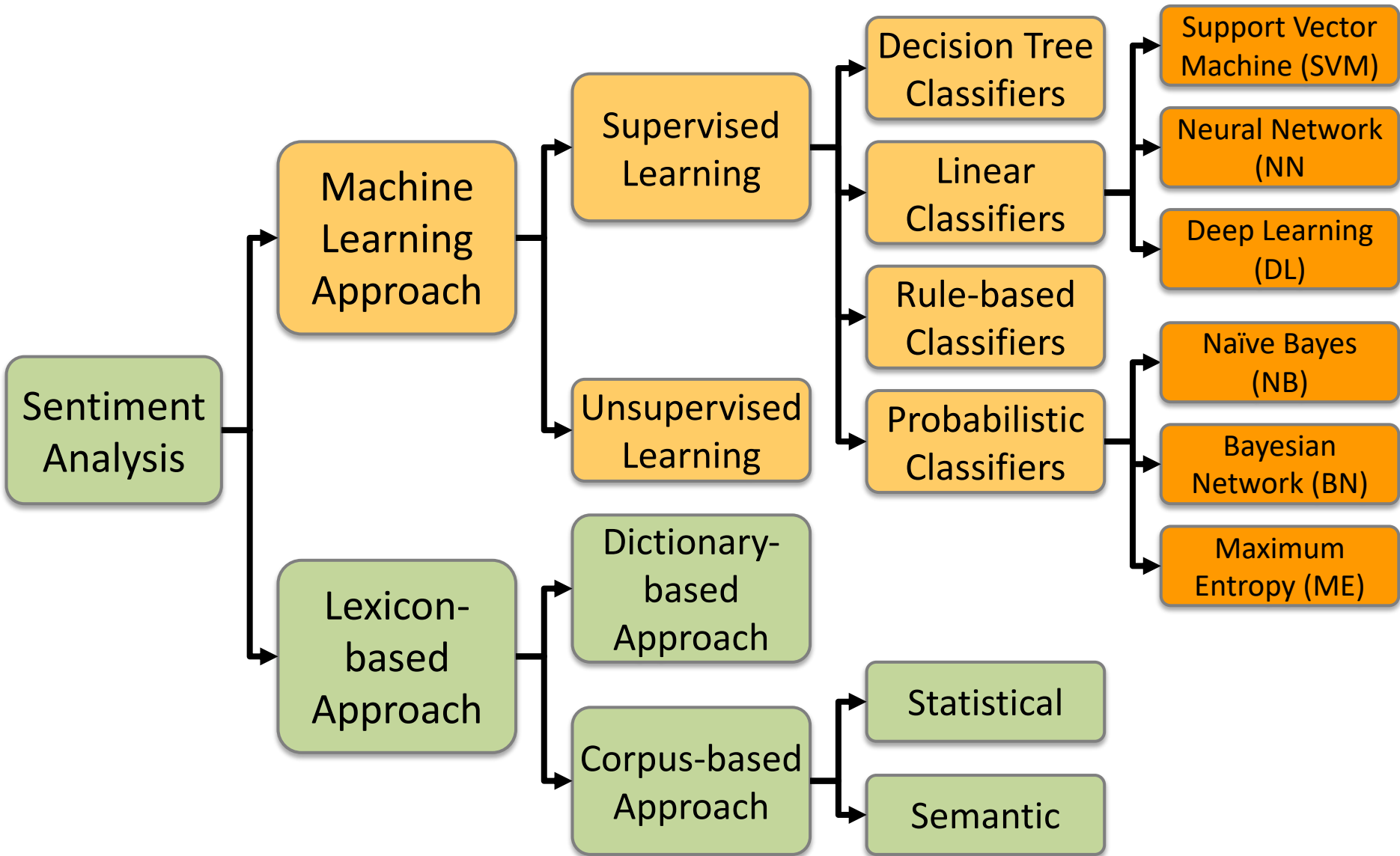
# A Multistep Process to Sentiment Analysis



# Sentiment Analysis



# Sentiment Classification Techniques





## Example of Opinion: review segment on iPhone



“I bought an iPhone a few days ago.

It was such a nice phone.

The touch screen was really cool.

The voice quality was clear too.

However, my mother was mad with me as I did not tell her before I bought it.

She also thought the phone was too expensive, and wanted me to return it to the shop. ... ”



# Example of Opinion: review segment on iPhone

“(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

(4) The voice quality was **clear** too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too expensive, and wanted me to return it to the shop. ...”

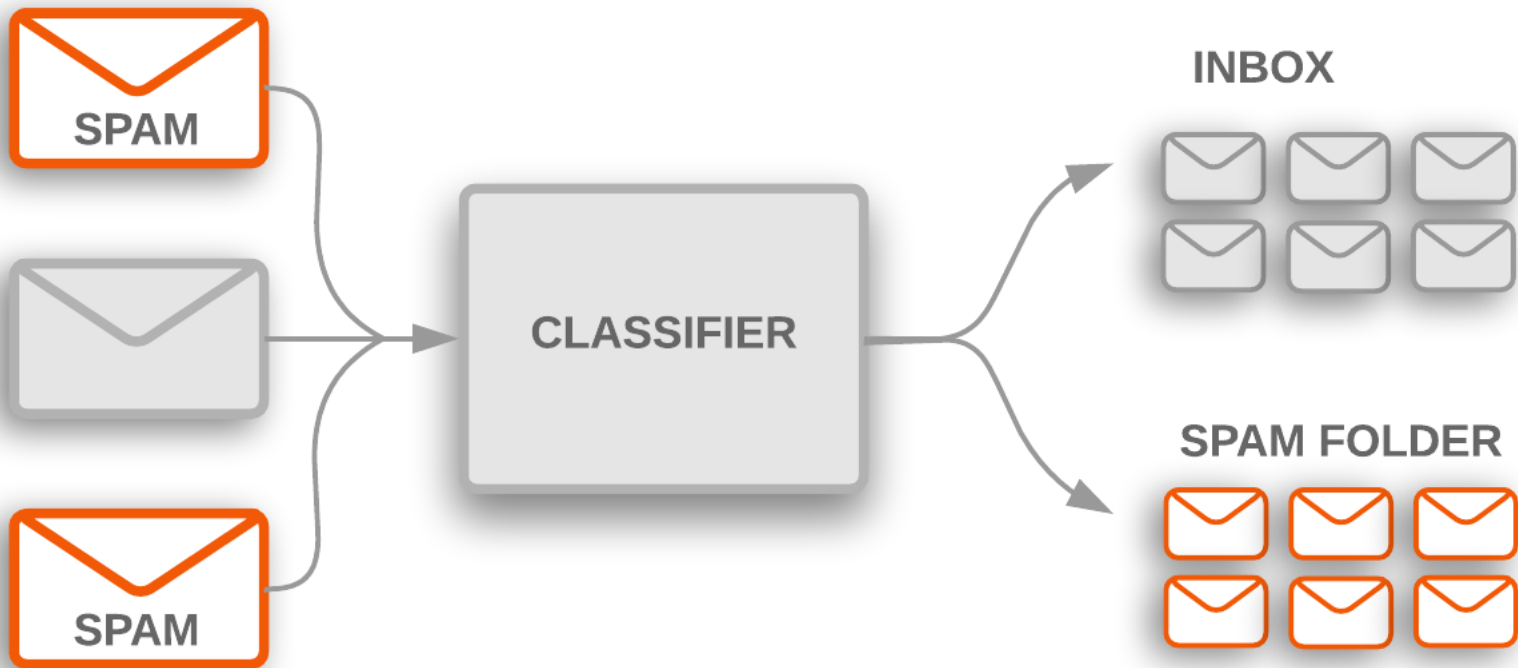


+Positive  
Opinion



-Negative  
Opinion

# Text Classification



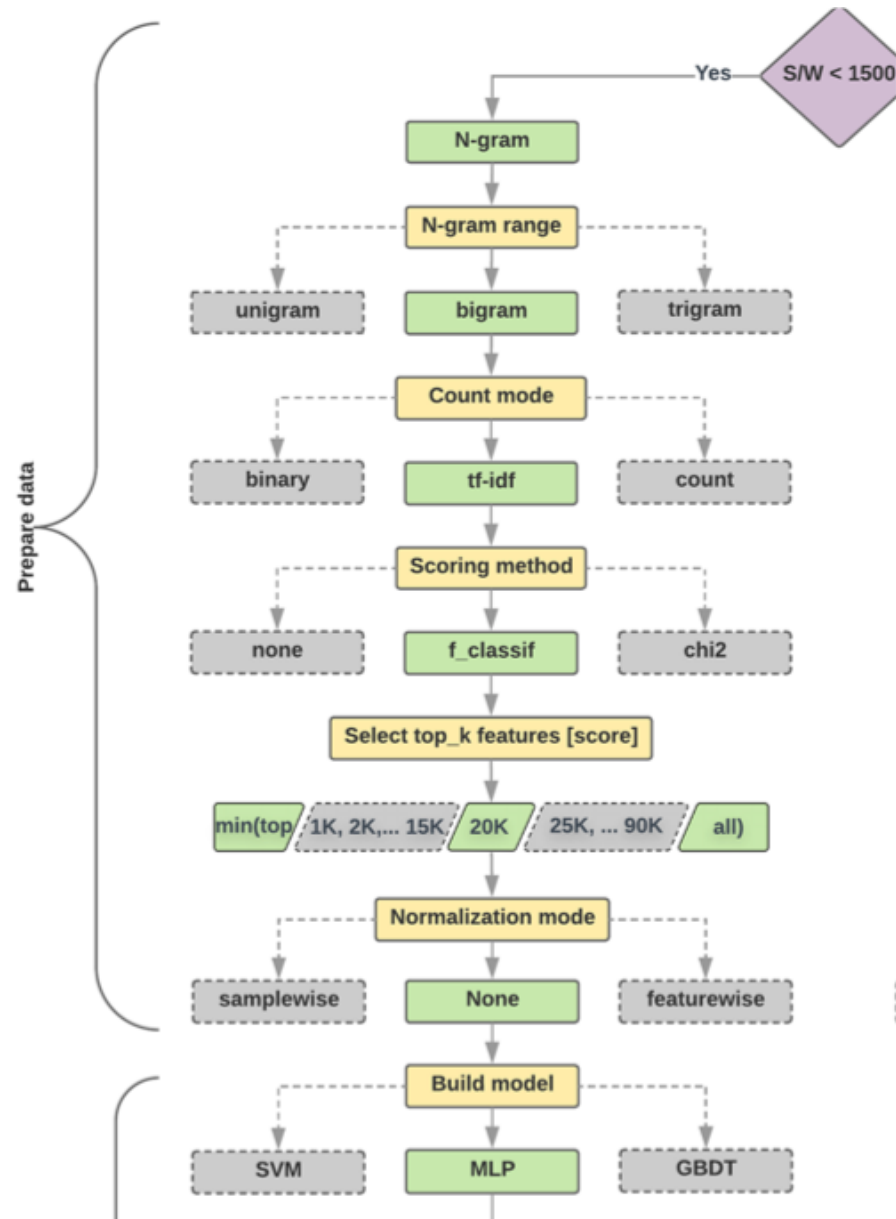
# Text Classification Workflow

- Step 1: Gather Data
- Step 2: Explore Your Data
- Step 2.5: Choose a Model\*
- Step 3: Prepare Your Data
- Step 4: Build, Train, and Evaluate Your Model
- Step 5: Tune Hyperparameters
- Step 6: Deploy Your Model

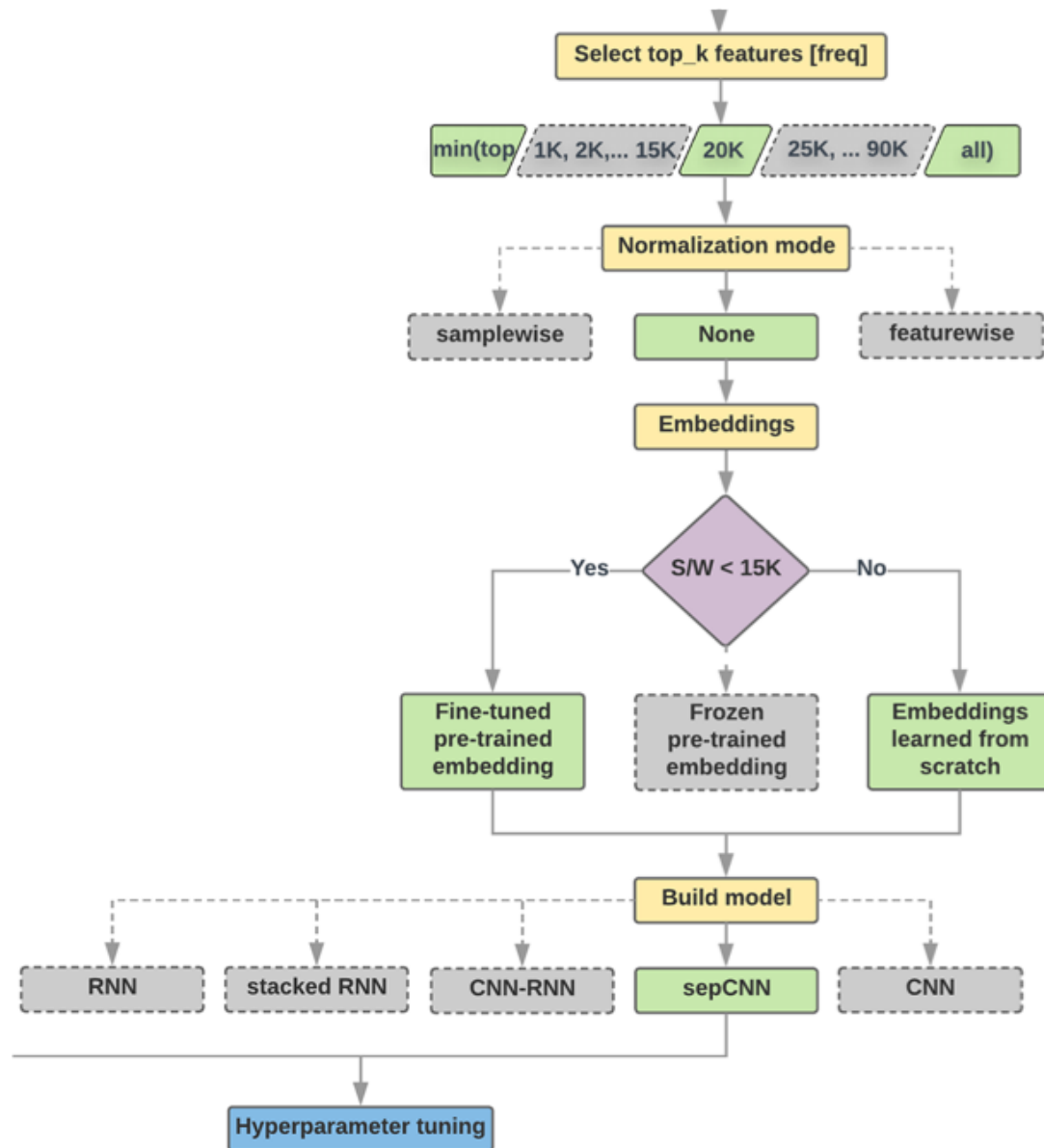




# Text Classification S/W<1500: N-gram



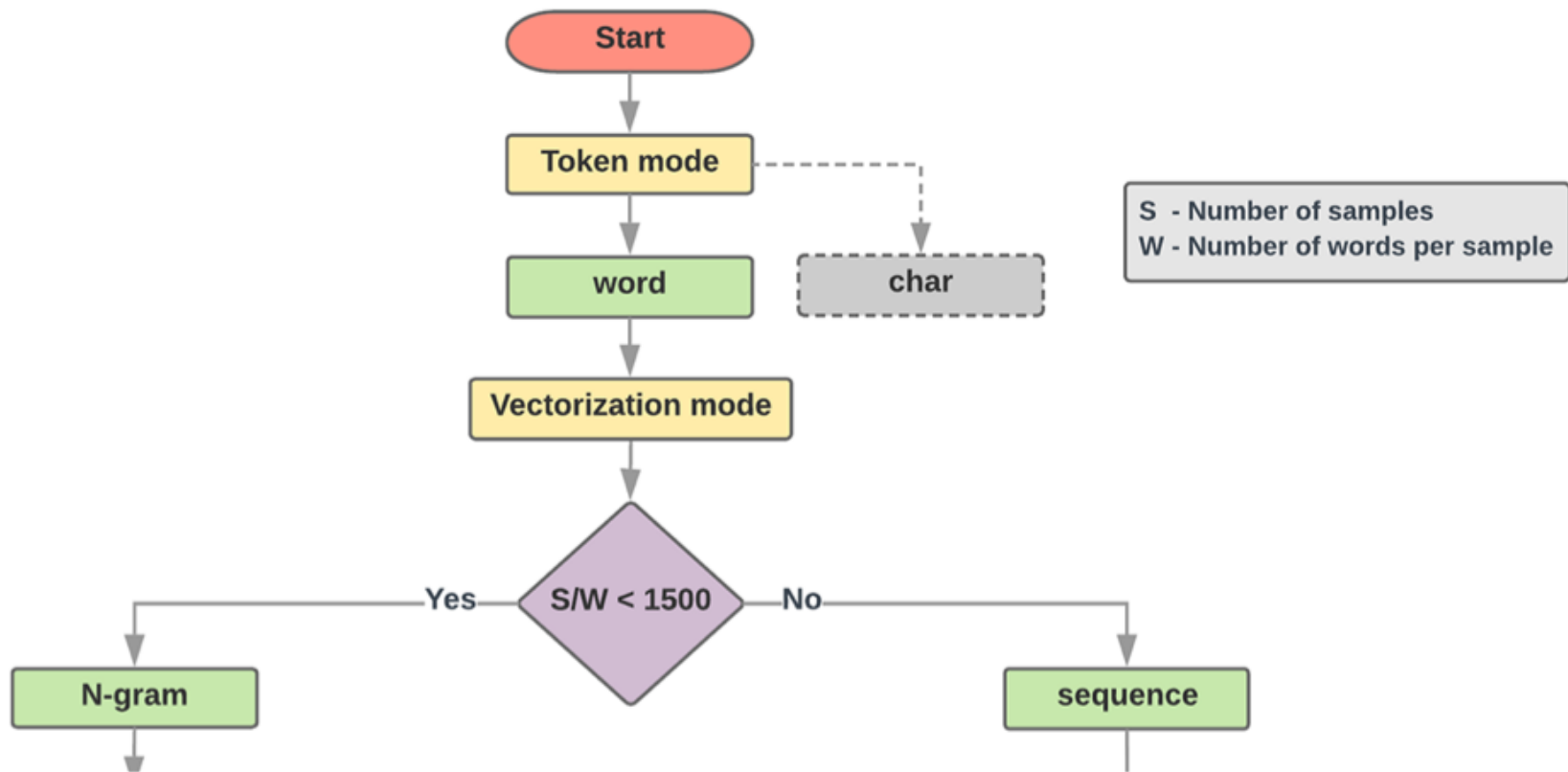
# Text Classification $S/W \geq 1500$ : Sequence



# Step 2.5: Choose a Model

**Samples/Words < 1500**

$$150,000/100 = 1500$$

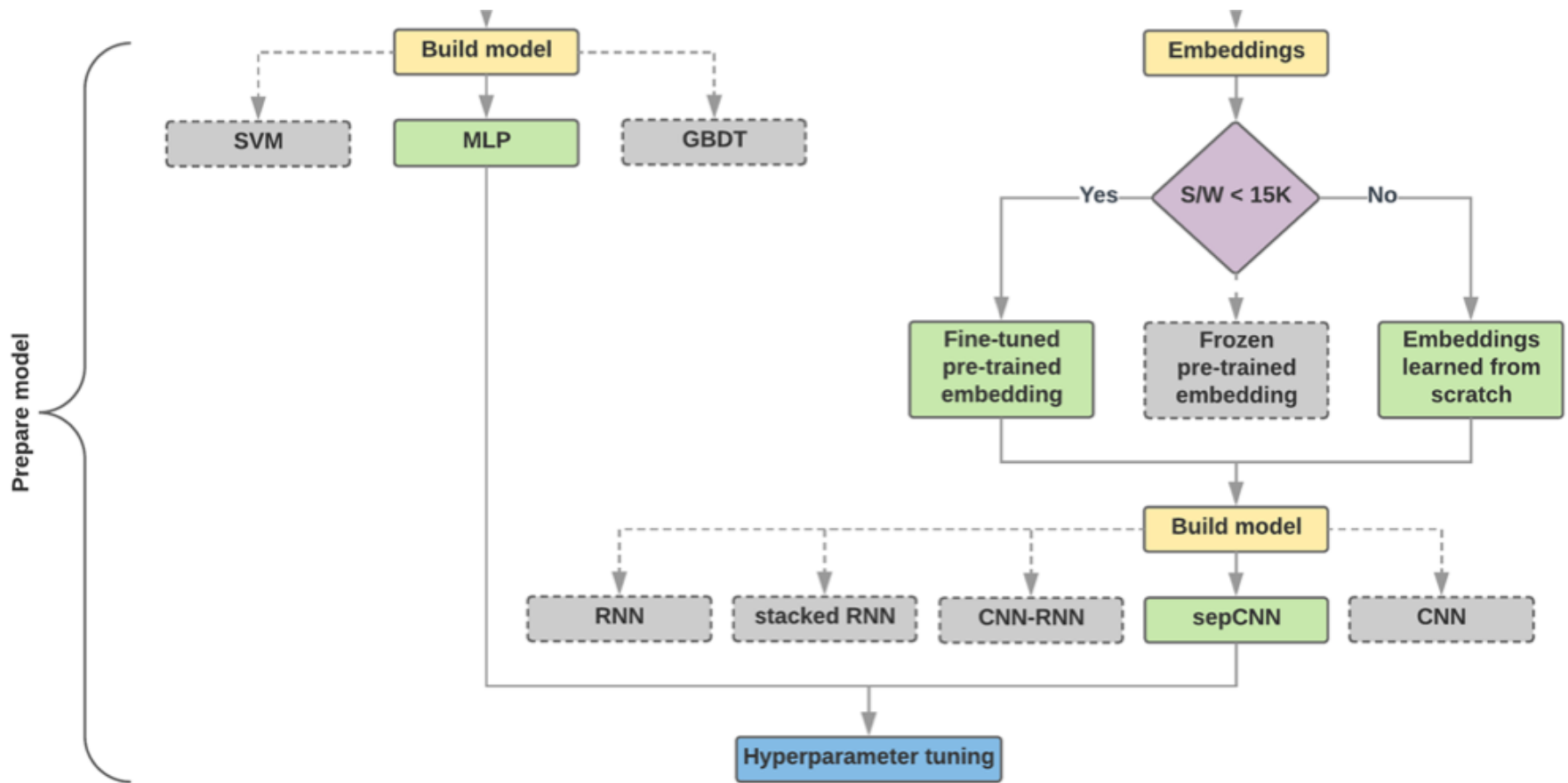


IMDb review dataset,  
the samples/words-per-sample ratio is ~ 144

# Step 2.5: Choose a Model

**Samples/Words < 15,000**

$$1,500,000/100 = 15,000$$





# Step 3: Prepare Your Data

Texts:

T1: 'The mouse ran up the clock'

T2: 'The mouse ran down'

Token Index:

```
{'the': 1, 'mouse': 2, 'ran': 3, 'up': 4, 'clock': 5, 'down': 6,}
```

NOTE: 'the' occurs most frequently,  
so the index value of 1 is assigned to it.  
Some libraries reserve index 0 for unknown tokens,  
as is the case here.

Sequence of token indexes:

T1: 'The mouse ran up the clock' =  
[1, 2, 3, 4, 1, 5]

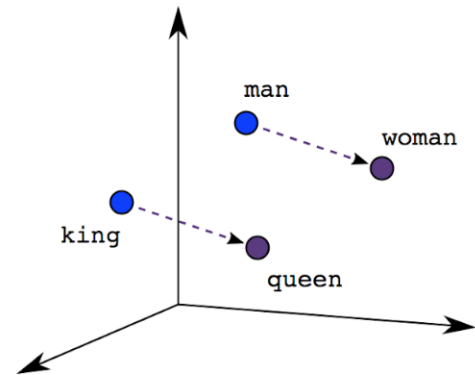
T2: 'The mouse ran down' =  
[1, 2, 3, 6]

# One-hot encoding

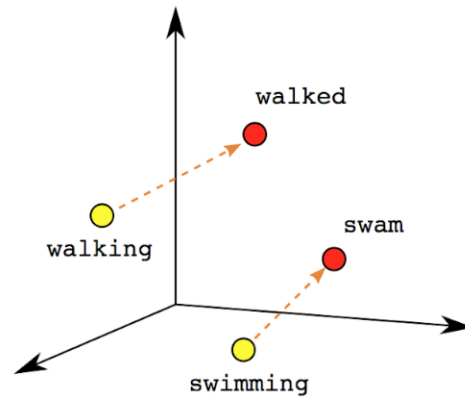
'The mouse ran up the clock' =

The	1	[	[0, 1, 0, 0, 0, 0, 0],
mouse	2		[0, 0, 1, 0, 0, 0, 0],
ran	3		[0, 0, 0, 1, 0, 0, 0],
up	4		[0, 0, 0, 0, 1, 0, 0],
the	1		[0, 1, 0, 0, 0, 0, 0],
clock	5		[0, 0, 0, 0, 0, 1, 0] ]
			[0, 1, 2, 3, 4, 5, 6]

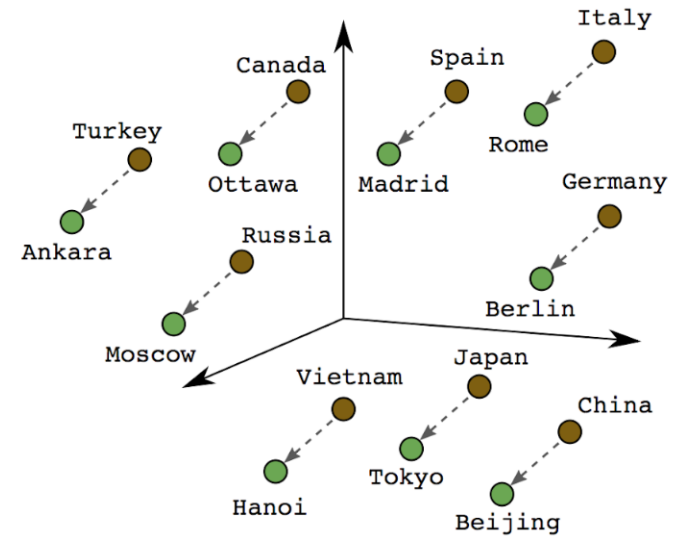
# Word embeddings



Male-Female

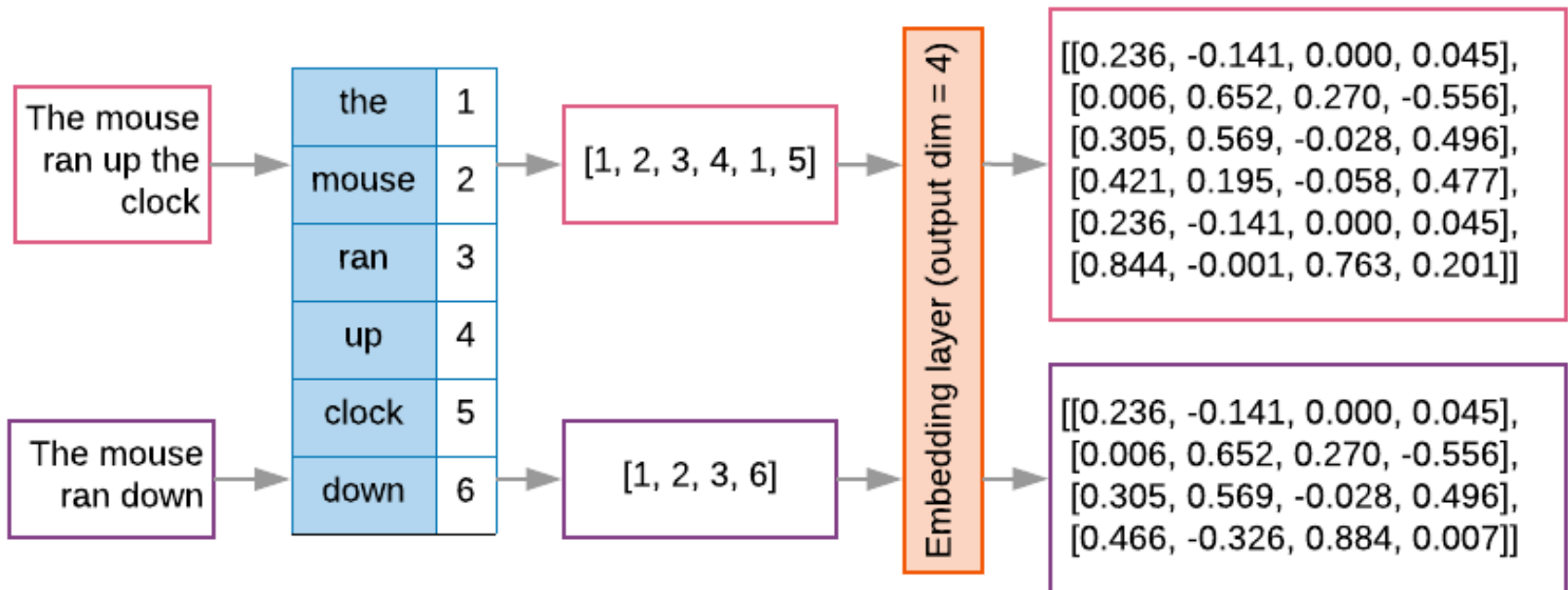


Verb Tense

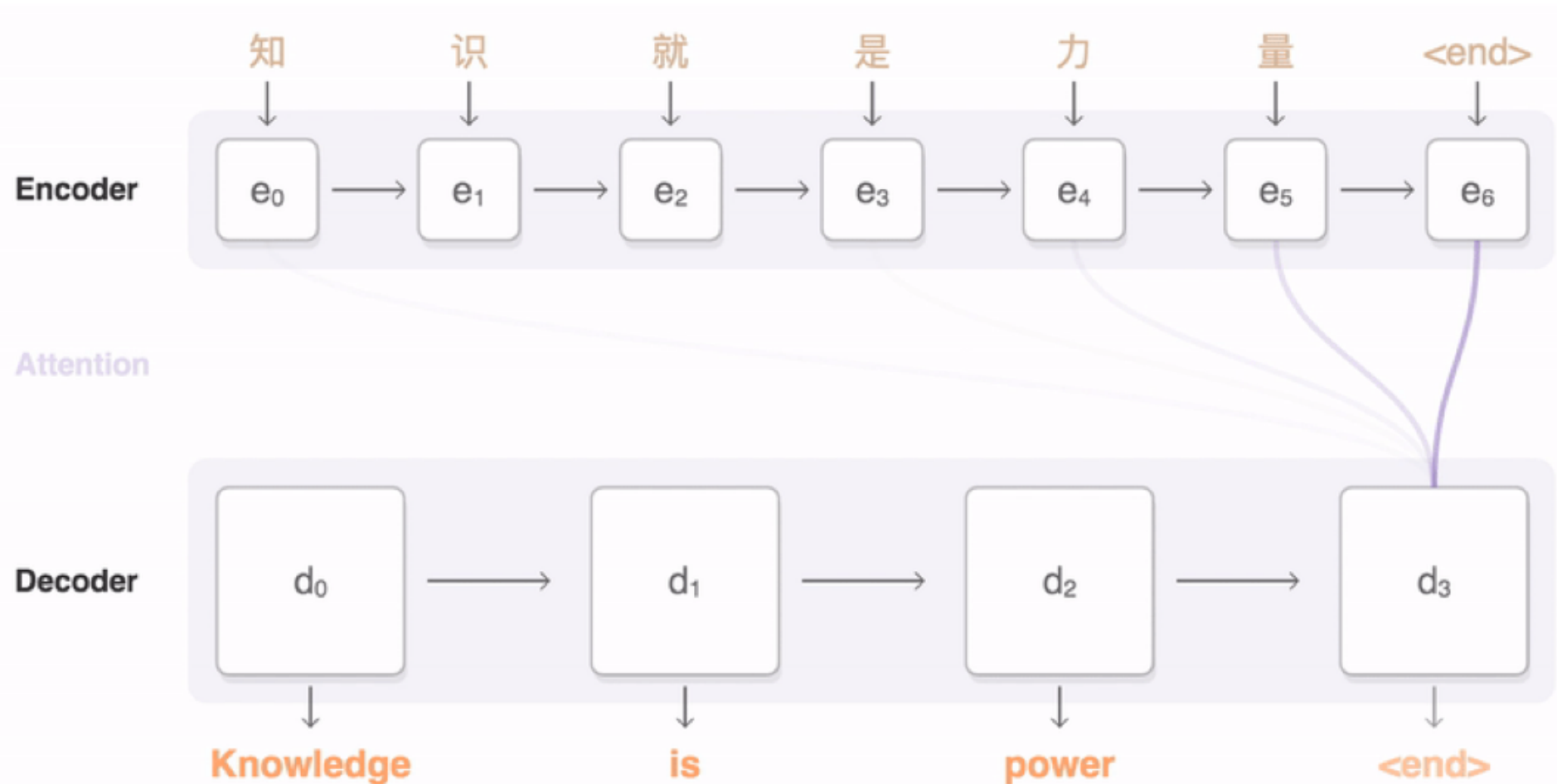


Country-Capital

# Word embeddings

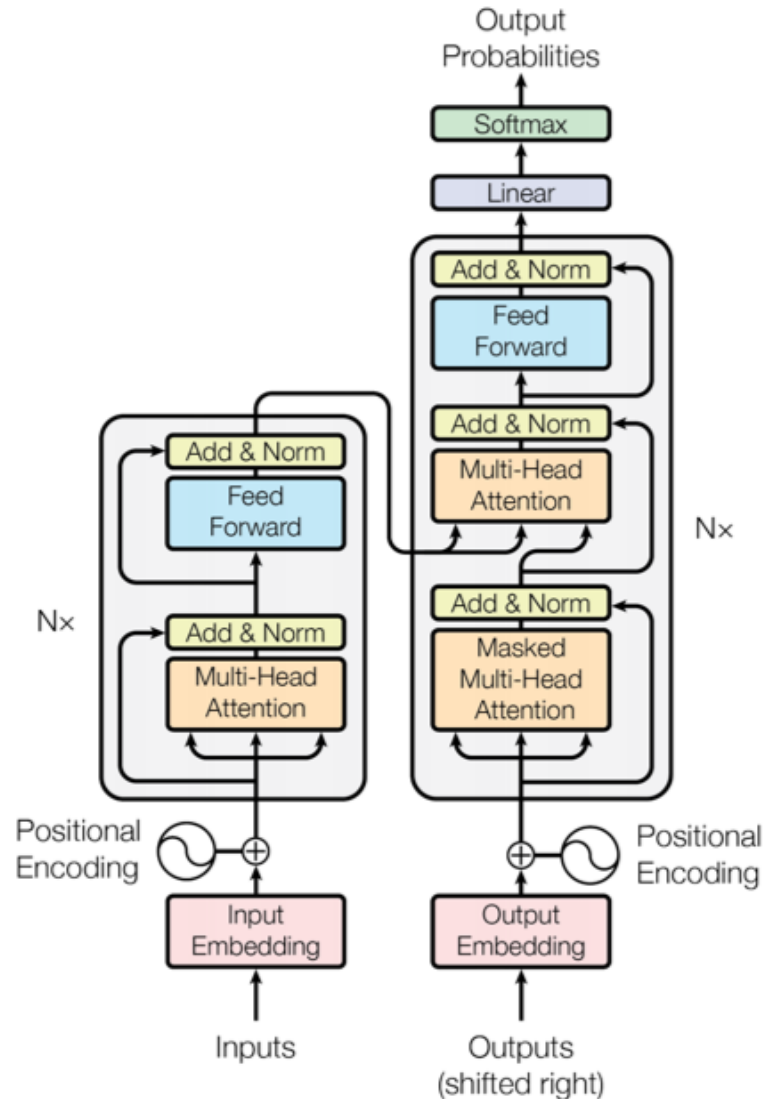


# Sequence to Sequence (Seq2Seq)



# Transformer (Attention is All You Need)

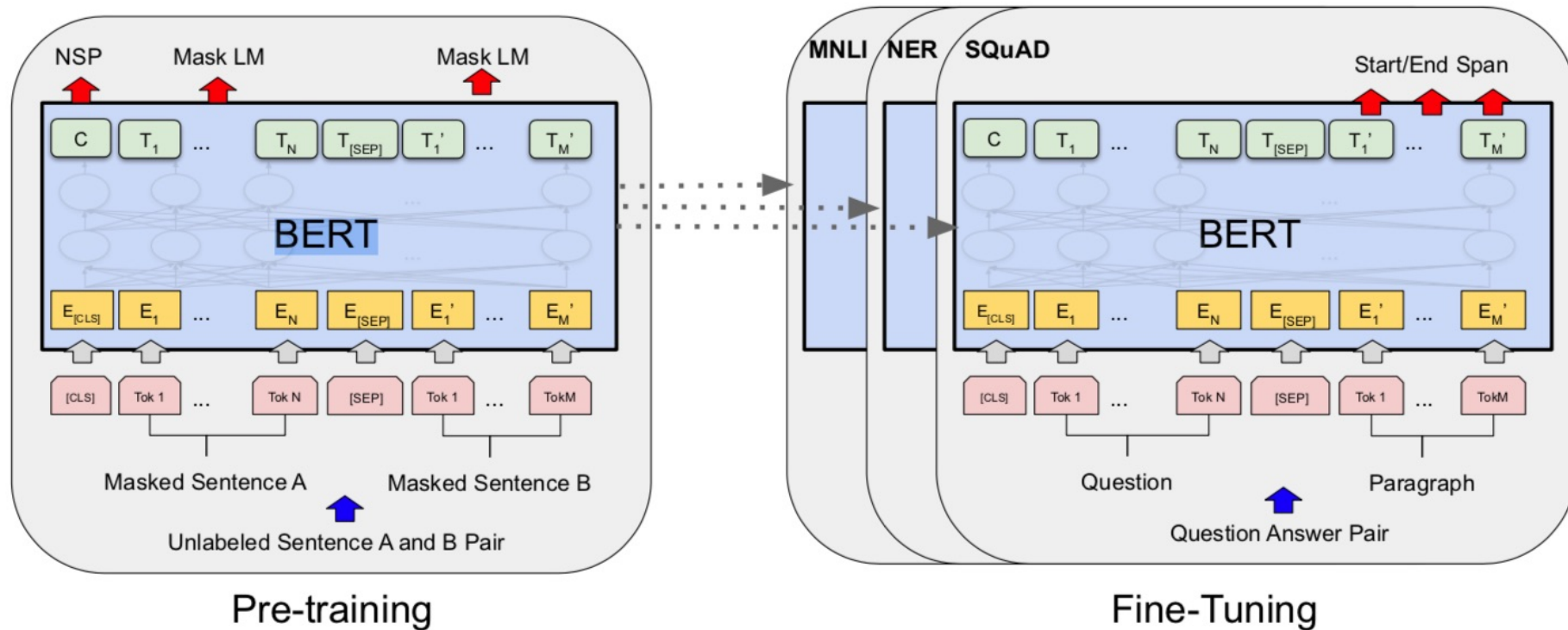
(Vaswani et al., 2017)



# BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

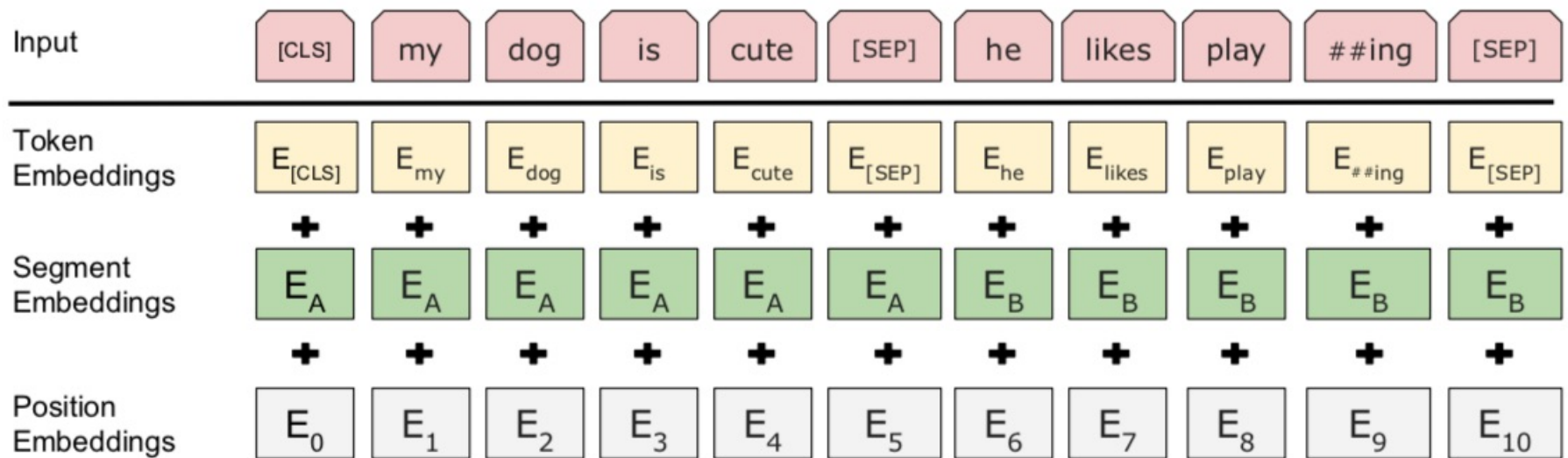
Overall pre-training and fine-tuning procedures for BERT



# BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

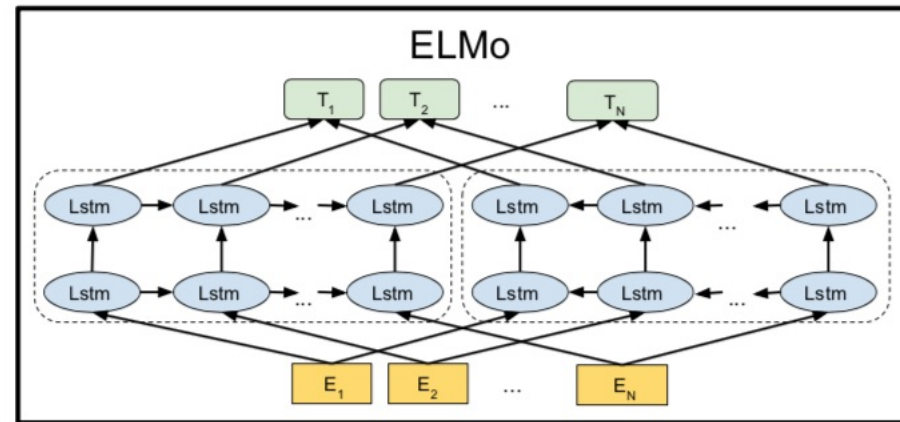
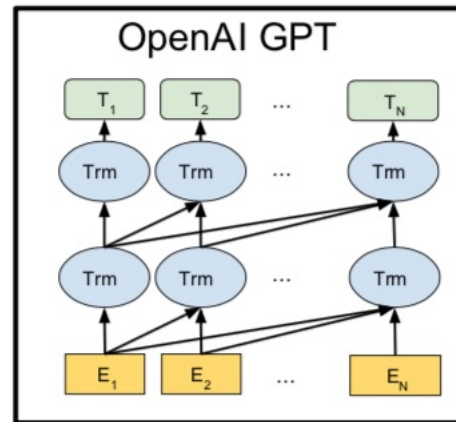
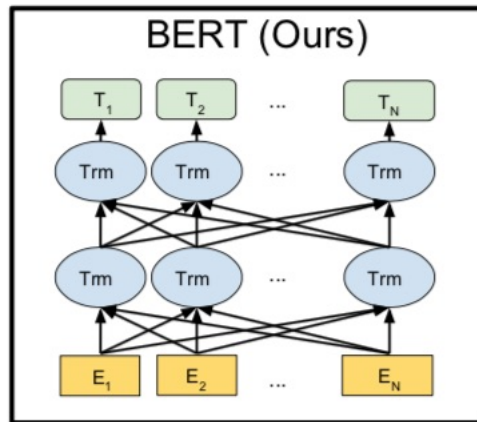
BERT (Bidirectional Encoder Representations from Transformers)

## BERT input representation

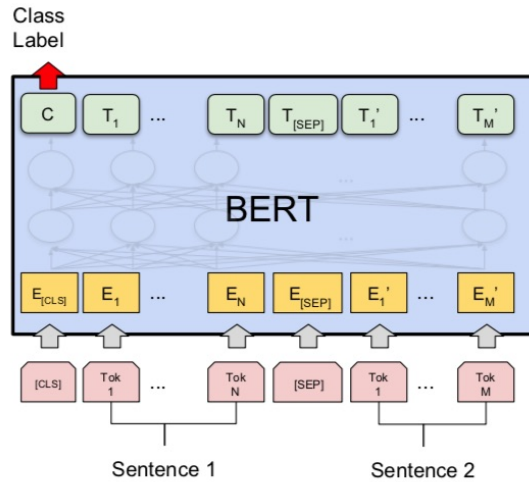




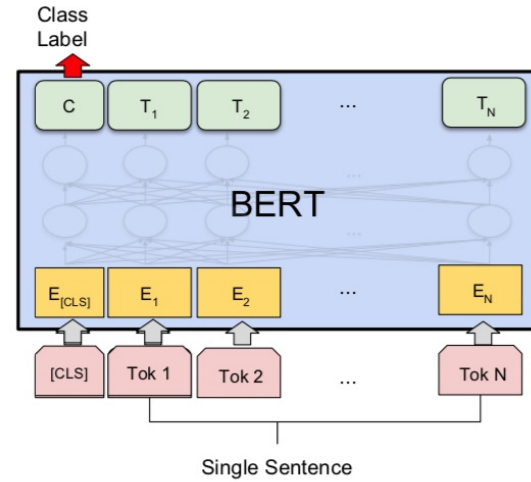
# BERT, OpenAI GPT, ELMo



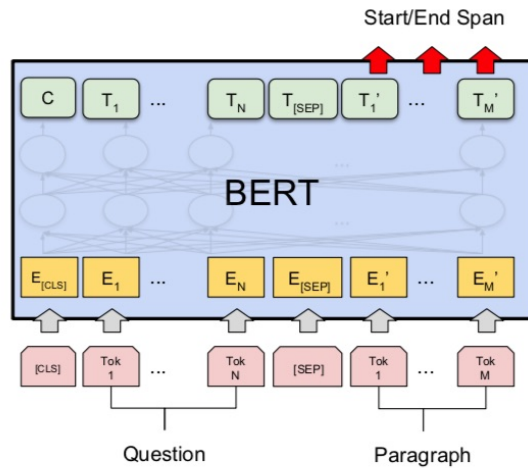
# Fine-tuning BERT on Different Tasks



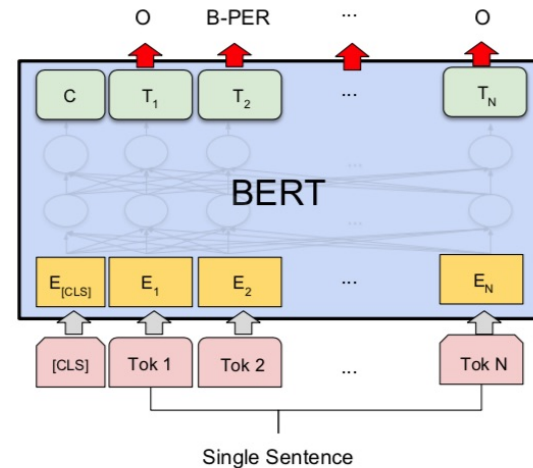
(a) Sentence Pair Classification Tasks:  
MNLI, QQP, QNLI, STS-B, MRPC,  
RTE, SWAG



(b) Single Sentence Classification Tasks:  
SST-2, CoLA



(c) Question Answering Tasks:  
SQuAD v1.1

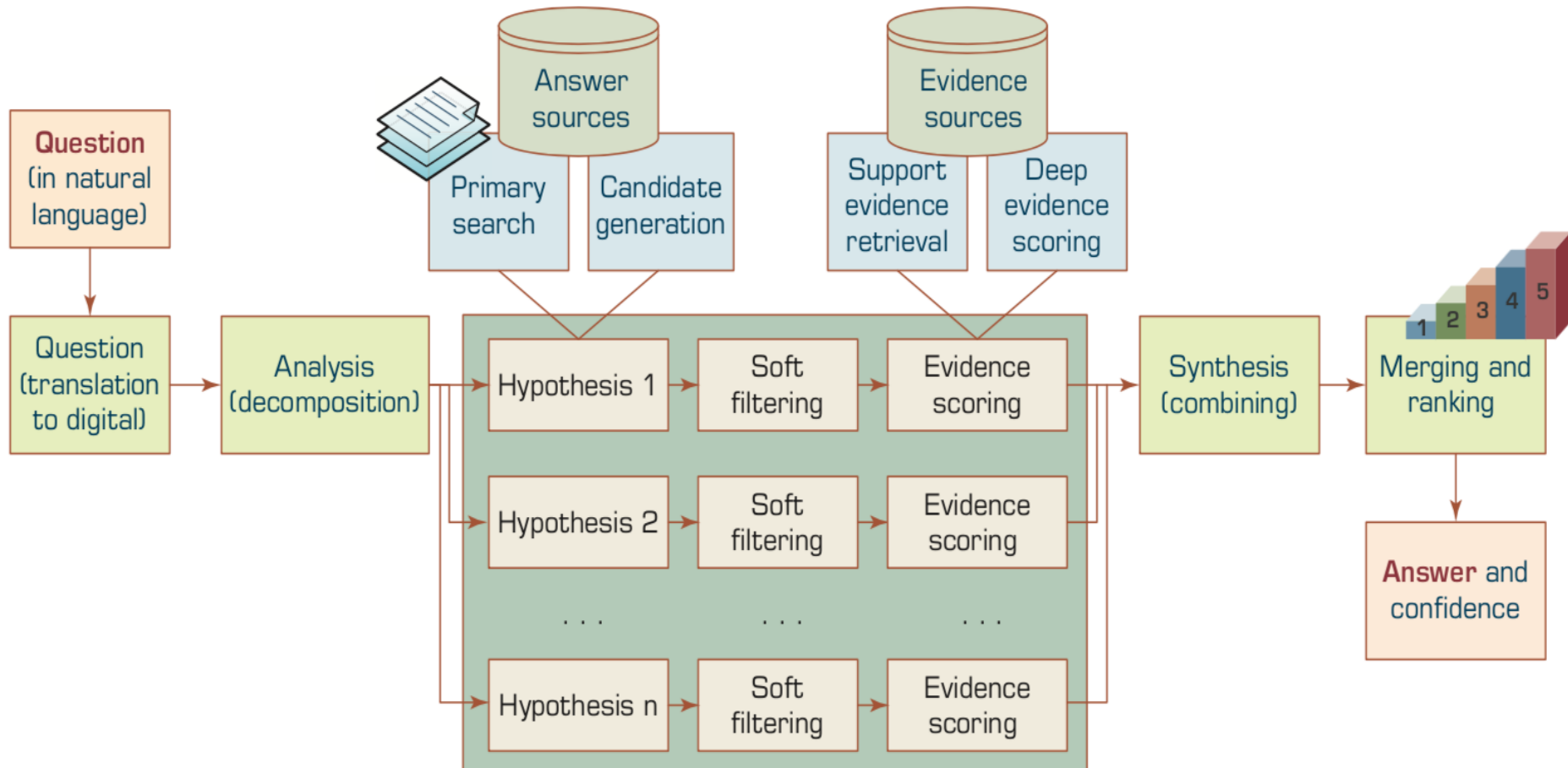


(d) Single Sentence Tagging Tasks:  
CoNLL-2003 NER

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).

"Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

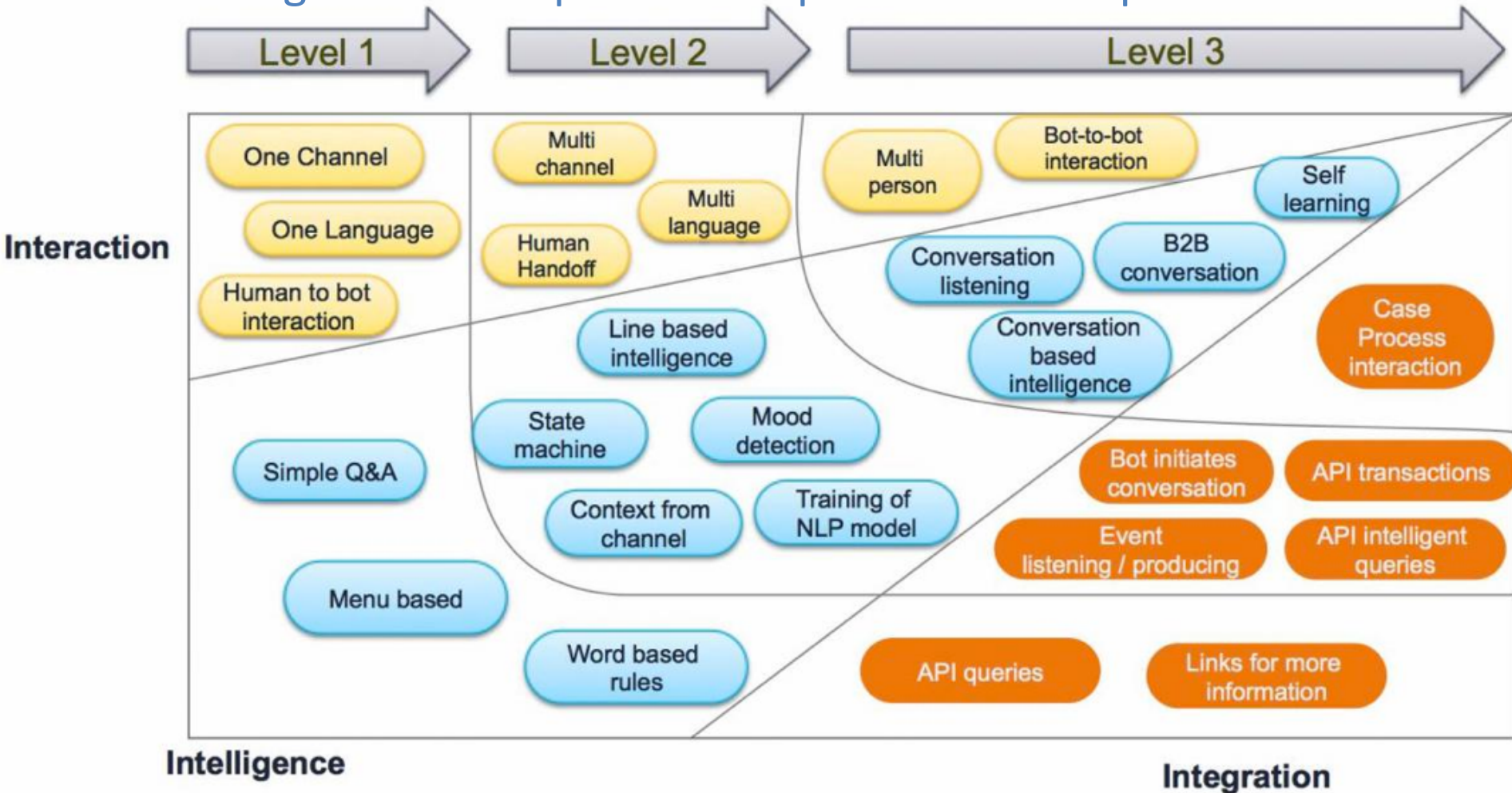
# A High-Level Depiction of DeepQA Architecture



# Chatbots

## Bot Maturity Model

Customers want to have simpler means to interact with businesses and get faster response to a question or complaint.



**Dialogue  
on  
Airline Travel  
Information System  
(ATIS)**

# The ATIS (Airline Travel Information System) Dataset

<https://www.kaggle.com/siddhadev/atis-dataset-from-ms-cntk>

<b>Sentence</b>	what	flights	leave	from	phoenix
<b>Slots</b>	O	O	O	O	B-fromloc
<b>Intent</b>	atis_flight				

Training samples: 4978

Testing samples: 893

Vocab size: 943

Slot count: 129

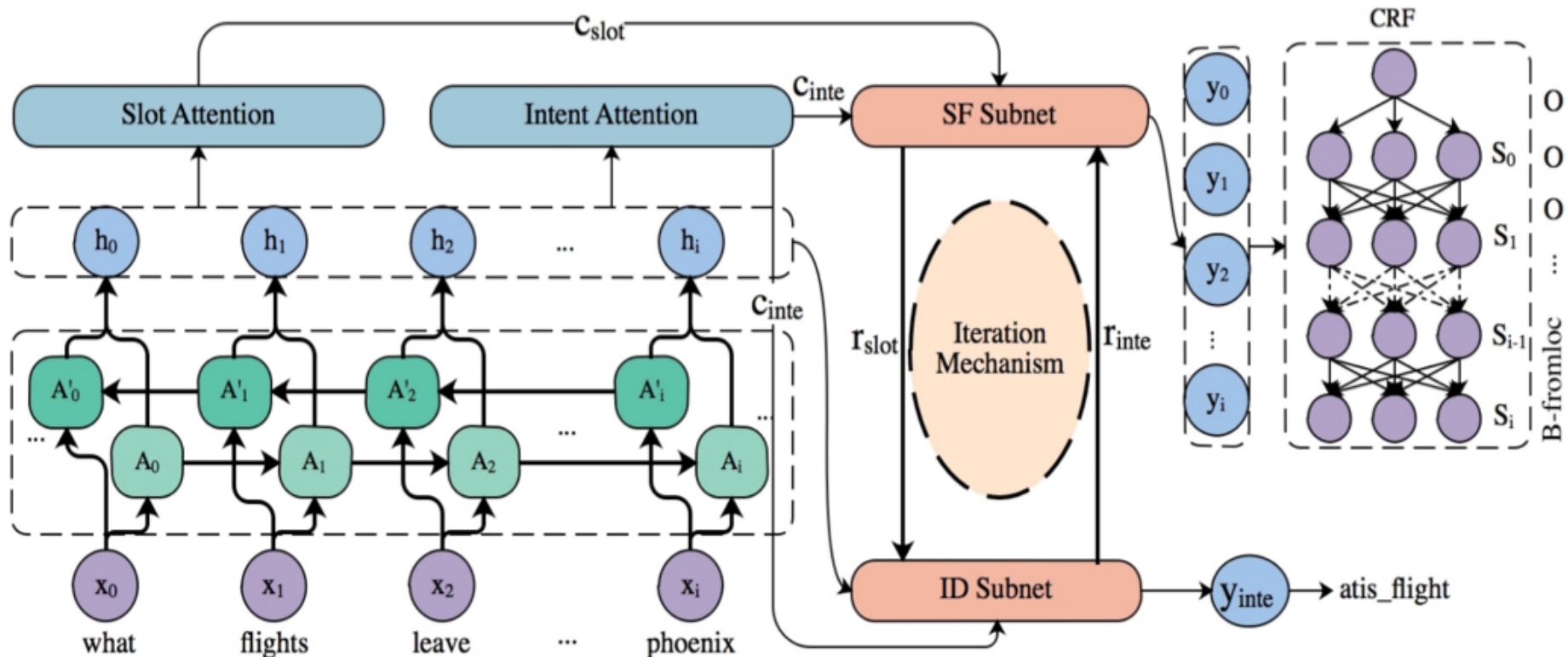
Intent count: 26

# SF-ID Network (E et al., 2019)

## Slot Filling (SF)

## Intent Detection (ID)

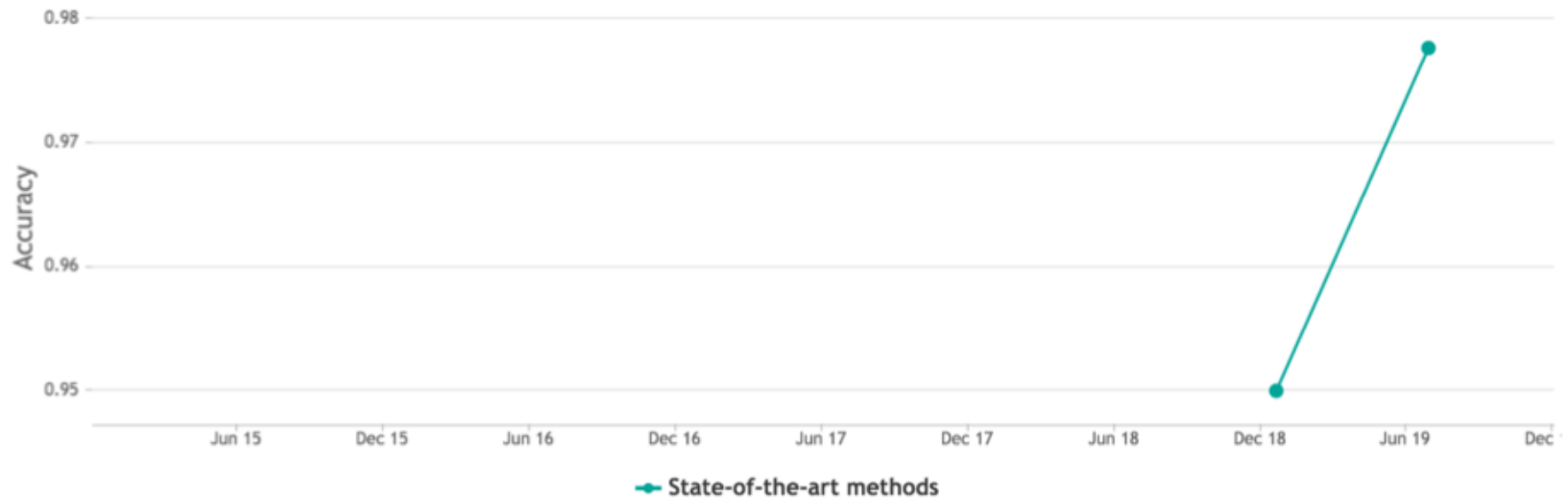
A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling



# Intent Detection on ATIS

## State-of-the-art

### Intent Detection on ATIS



[Edit](#)

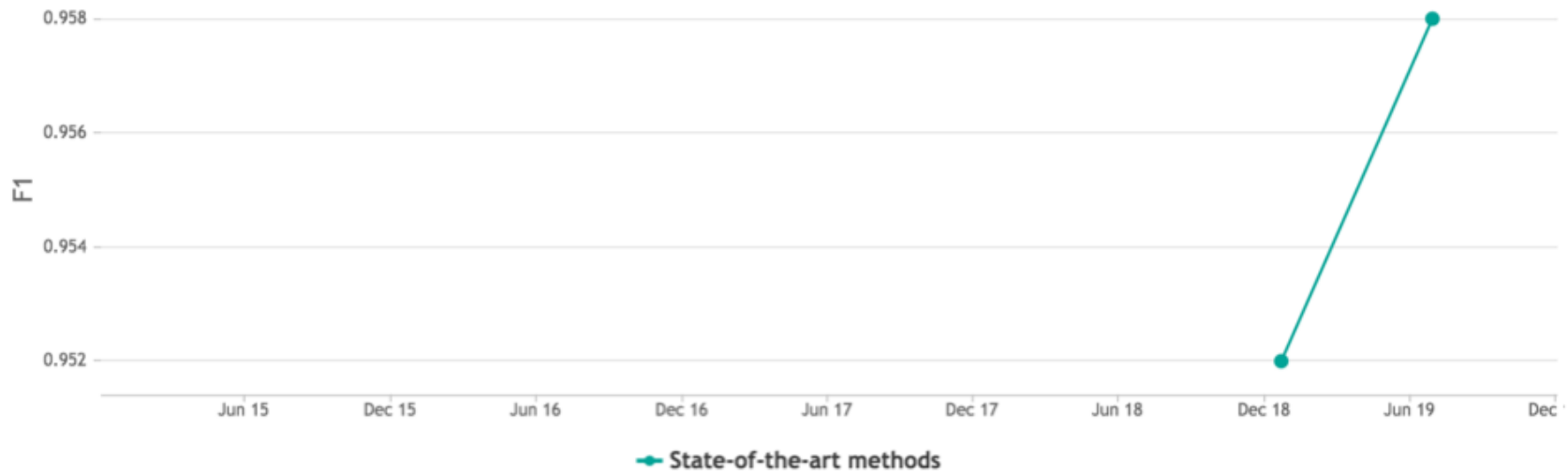
RANK	METHOD	ACCURACY	PAPER TITLE	YEAR	PAPER	CODE
1	SF-ID	0.9776	<a href="#">A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling</a>	2019		
2	Capsule-NLU	0.950	<a href="#">Joint Slot Filling and Intent Detection via Capsule Neural Networks</a>	2018		



# Slot Filling on ATIS

## State-of-the-art

### Slot Filling on ATIS



Edit

RANK	METHOD	F1	PAPER TITLE	YEAR	PAPER	CODE
1	SF-ID	0.958	<a href="#">A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling</a>	2019		
2	Capsule-NLU	0.952	<a href="#">Joint Slot Filling and Intent Detection via Capsule Neural Networks</a>	2018		

Source: <https://paperswithcode.com/sota/slot-filling-on-atis>

# Restaurants Dialogue Datasets

- MIT Restaurant Corpus
  - <https://groups.csail.mit.edu/sls/downloads/restaurant/>
- CamRest676  
(Cambridge restaurant dialogue domain dataset)
  - <https://www.repository.cam.ac.uk/handle/1810/260970>
- DSTC2 (Dialog State Tracking Challenge 2 & 3)
  - <http://camdial.org/~mh521/dstc/>

# 任務型對話系統

## The Evaluation of Chinese Human-Computer Dialogue Technology, SMP2019-ECDT

- 自然語言理解  
Natural Language Understanding (NLU)
- 對話管理  
Dialog Management (DM)
- 自然語言生成  
Natural Language Generation (NLG)

# Summary

- This course introduces the **fundamental concepts and research issues of Text Mining.**
- Topics include
  - Foundations of Text Mining: Natural Language Processing (NLP),
  - Python for NLP,
  - Processing and Understanding Text,
  - Feature Engineering for Text Representation,
  - Text Classification,
  - Text Summarization and Topic Models,
  - Text Similarity and Clustering,
  - Semantic Analysis and Named Entity Recognition,
  - Sentiment Analysis,
  - The Promise of Deep Learning and Universal Sentence-Embedding Models,
  - Question Answering and Dialogue Systems,
  - and Case Study on Text Mining.

# 文字探勘 (Text Mining)

## Contact Information



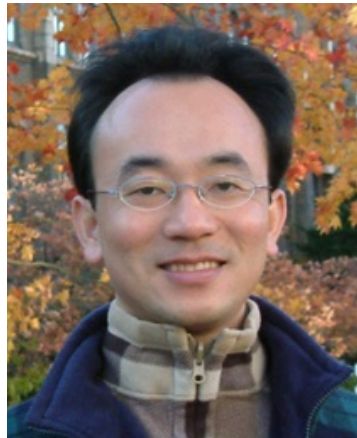
Chichang Jou

周清江

Associate Professor

副教授

[cjou@mail.tku.edu.tw](mailto:cjou@mail.tku.edu.tw)



Min-Yuh Day

戴敏育

Associate Professor

副教授

[myday@mail.tku.edu.tw](mailto:myday@mail.tku.edu.tw)

淡江大學 資訊管理學系

Department of Information Management, Tamkang University

