NLP

Natural Language Processing

概念

- 即人類語言 (文本符號/語音信號)，思維的載體和交流的工具。
- 電腦科學與語言學的交叉學科，屬AI的一個重要分支，亦稱電腦語言學 (Computational Linguistics, CL)。

運算智能 → 感知智能 → 認知智能

- 運算智能：電腦基礎運算和儲存能力
- 感知智能：電腦的模式識別能力(語音識別或圖像識別)
- 認知智能：涉及自然語言處理及常識建模和推理等研究

難點

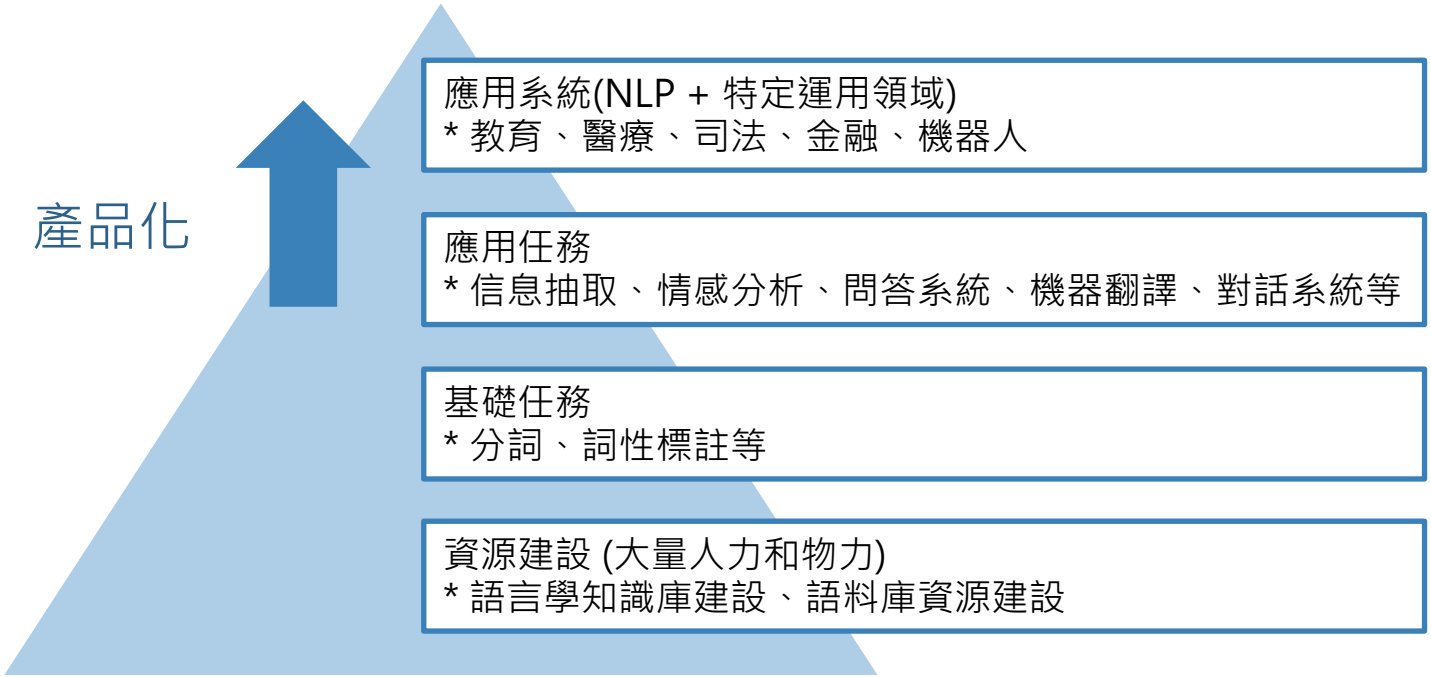
- 抽象性：車-汽車、火車、腳踏車。
- 語義組合性 ：有限符號可組成無限語義。
- 歧義性：一詞多義-如蘋果。

  形式不同語意相同-阿湯哥演捍衛戰士/捍衛戰士男主角是湯姆克魯斯。
- 進化性：新詞彙層出不窮- 新冠 / 舊詞彙有新含義 – 杯具。
- 非規範性：音近詞 (484→是不是)。

  單詞簡寫或變形(please → pls、cool →cooooooool)。

  錯別字……。
- 理解語言需背景知識和推理能力等。

任務層級


產品化
應用系統(NLP + 特定運用領域)
* 教育、醫療、司法、金融、機器人
應用任務
* 信息抽取、情感分析、問答系統、機器翻譯、對話系統等
基礎任務
* 分詞、詞性標註等
資源建設 (大量人力和物力)
* 語言學知識庫建設、語料庫資源建設

發展歷史

小規模專家知識
20世紀50年代 – 90年代

大規模語料庫
統計模型
20世紀90年代 – 21世紀初

- 特徵工程

大規模語料庫
深度學習
2010年 – 2017年

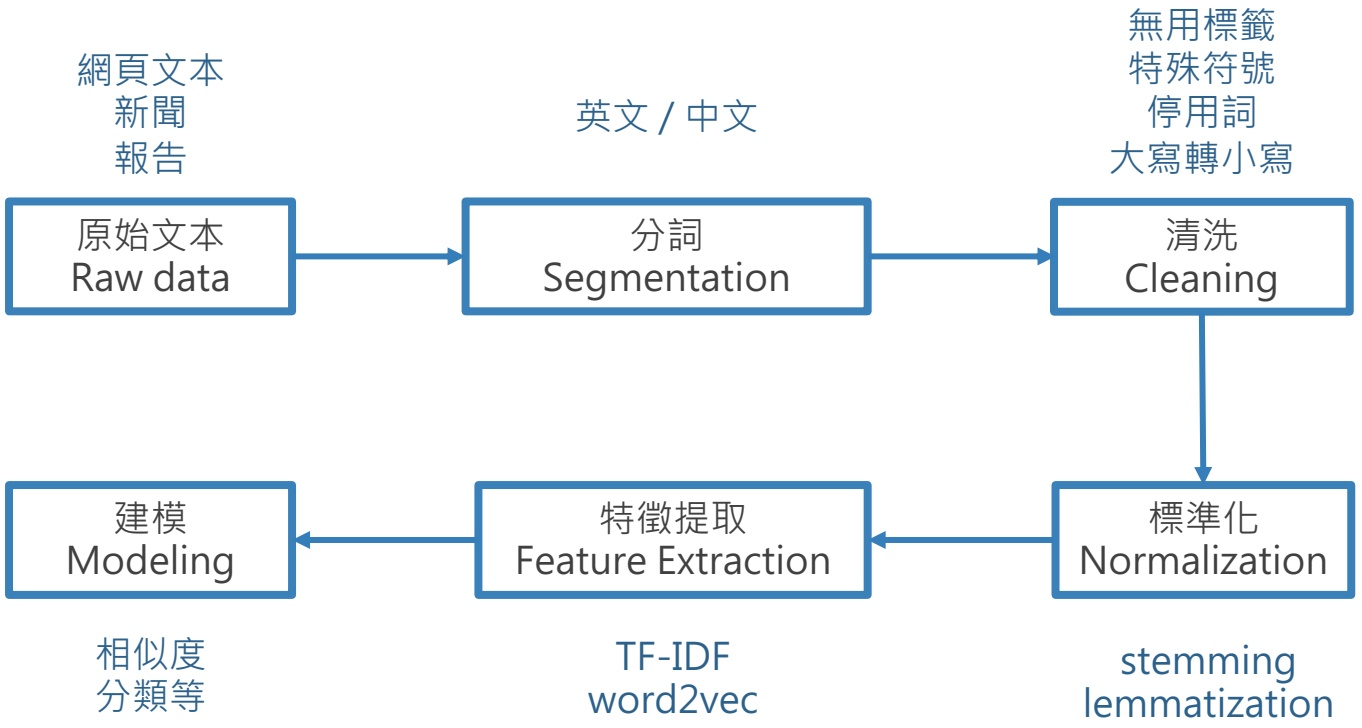
- 自動發現有效特徵
- 跨任務、跨語言、跨模態遷移

大規模預訓練
語言模型
2018年 – 至今

- 模型預訓練（Pre-train），在一個原任務上預先訓練初始模型，然後在下游任務上繼續對該模型進行精調（Fine-tune），從而達到提高下游任務準確率的目的

自然語言處理流程

Pipeline


網頁文本
新聞
報告
英文 / 中文
無用標籤
特殊符號
停用詞
大寫轉小寫
原始文本
Raw data
分詞
Segmentation
清洗
Cleaning
標準化
Normalization
特徵提取
Feature Extraction
建模
Modeling
相似度
分類等
TF-IDF
word2vec
stemming
lemmatization

分詞

- 中文 – jieba (結巴)
- 英文 – 空格

清洗

特徵篩選

- Filtering Words：把**停用詞**和**出現頻率很低**的詞過濾

Removing Stop Words (考慮應用場景)

- 英文：如the, an, I, Wow
- 中文：如「好」,「很好」(語意分析就不適合過濾)

標準化 – 英文

多運用在非中文的語言
如英文語句中，同一個單詞可能隨著時態、單複數、主被動等狀況不同，ex: running 與 run。
Stemming 與 Lemmatization的目的就是及將這些不同的表示型態標準化，以降低文本複雜度。

- Stemming (詞幹提取): 把文字的後面整個切掉，而不在意切掉後的字是不是字典上有的字。偏rule-base的方式拆解（語言學）

University, universal , universities , universe 返回 univers

- Lemmatization (詞形還原): 盡可能把恢復成字典上有的字

Amused, amusing

stemming 返回 amus，Lemmatization 返回 amuse

標準化 - Porter Stemmer

Step 1a

sses	→ ss	caresses	→ caress
ies	→ i	ponies	→ poni
ss	→ ss	caress	→ caress
s	→ ø	cats	→ cat

Step 1b

(*v*)ing	→ ø	walking	→ walk
		sing	→ sing
(*v*)ed	→ ø	plastered	→ plaster

Step 2 (for long stems)

ational	→ ate	relational	→ relate
izer	→ ize	digitizer	→ digitize
ator	→ ate	operator	→ operate
…			

Step 3 (for longer stems)

al	→ ø	revival	→ reviv
able	→ ø	adjustable	→ adjust
ate	→ ø	activate	→ activ

標準化 – 重要說明

- Stemming (詞幹提取)和Lemmatization (詞形還原)要視情況使用，因會縮減文本傳達的信息內容和意義。

- 如應用中包含搜索過程，上述詞彙壓縮法會導致搜尋引擎返回更多與詞的原意不相關的文檔。

特徵提取 - TF-IDF

TF (Term Frequency) 詞頻

每個詞在每個文件出現的比率

如一篇文件中，被我們篩選出兩個重要名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 70 次，「富有」出現 30 次，那「健康」的 tf = 70 / (70+30) = 70/100 = 0.7，「富有」的 tf = 30 / (70+30) = 30/100 = 0.3

IDF (Inverse Document Frequency) 逆向檔案頻率

詞在所有文件的頻率，頻率越高表該詞越不具代表性，IDF值越小

『文章數總和』除以『該字詞出現過的文章篇數』後，取log值

$$idf_t = \log\left(\frac{D}{d_t}\right)$$

如有100個網頁，「健康」出現在 10 個網頁當中，而「富有」出現在 100 個網頁當中，那麼「健康」的 idf=log(100/10)=1，而「富有」的idf=log(100/100)=0。所以，「健康」出現的機會小，與出現機會很大的「富有」比較起來，便顯得非常重要。

TF-IDF

TF*IDF(出現頻率*代表性)，篩選重要關鍵字

Thanks!