Digital Speech Recognition Final Report

R08922160
山下 夏輝 (Yamashita Natsuki)

《My Background 簡介》

慶應義塾大學 文學部 圖書館資訊學系 畢業。
之後，當日文老師、日文教學教材編輯負責人，多文化多語言人員的 Project Management Office 負責人。
透過這些經驗，獲得的心得如下

1. 不管是表達能力、語言能力多好，如果使用的 communication approach 不適合的話，會產生問題。
2. 如果機器翻譯的精準度、正確性提高了的話，語言教學的價值可能會降低。反而跨文化跨語言時的 communication approach、技巧、功夫的教學的價值可能會提升。

有了上述的經驗和心得了之後，因為我想找到需要多文化多語言時能夠幫助到 communication 的技術、知識、觀點，所以就學國立台灣大學學習最新的科技。

《My purpose of this final project》

根據上述的自己興趣，以課堂上學到的東西為基礎、閱讀論文、知道更多技術應用和方式、發想對自己的興趣可以應用的。

具體構成和方法如下：

1. 《What I check mainly》
    a. 跟這堂課相關並且我有興趣的 query 來查詢論文
    b. 閱讀自己有興趣的論文
    c. 簡單的總結閱讀的論文
2. 《What I studied》
    a. 不懂的 technical word 調查學習
    b. 在這 report 上簡單的總結
    c. 在這 report 上記小小的發現，類似課堂上沒有印象深刻或是第一次知道了的事請
3. 《What I think & How I feel》
    a. 心得
4. 《What is my idea》
    a. 可以去研究的題目 idea list

《What I check mainly》

//Query 用「」來表示，用 google sholar 查詢。
//小小的發現，用 Note：以右記述
//這裡有的是我閱讀的論文的簡單的 summary，

「音声認識」
「音声言語情報」

高津弘明, et al. 会話によるニュース記事伝達のための音声合成. 人工知能学会論文誌, 2019, 34.2: B-I65_1-15.

課題：機器把新聞告訴人的時候，停頓對聽眾的理解會有什麼樣的效果呢？聞內容理解成度，導出實驗結果。在 speech generation 當中適當的停頓對聽眾來說容不容易理解。

結果：機器說話時，對聽者的理解程度和容易問問題的氣氛，有調整停好的頓有正的效果

Train 停頓時，用 sentence 比用 paragraph 效果比較好

技術：LSTM，DNN

Model：

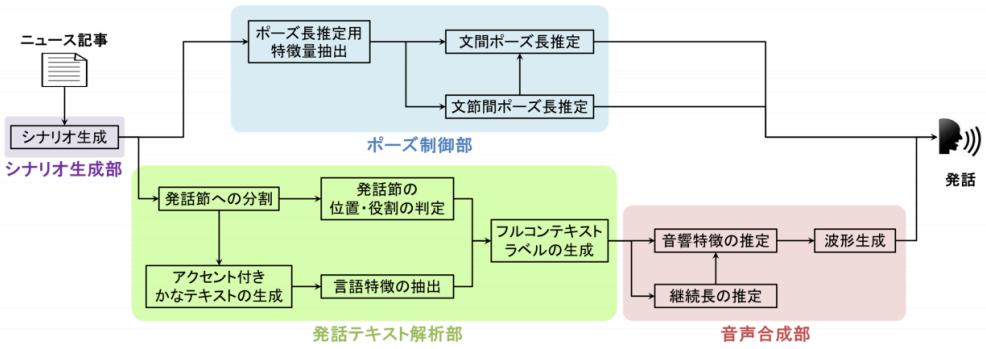


図 1　システム構成図

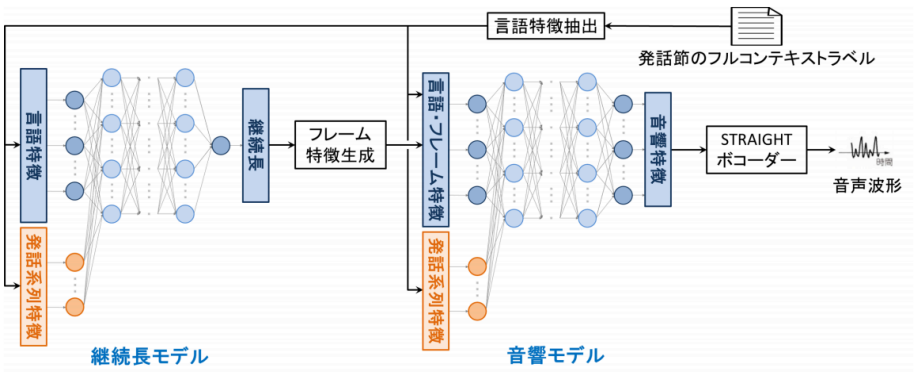


図 9　発話系列特徴を考慮した DNN 音声合成モデル

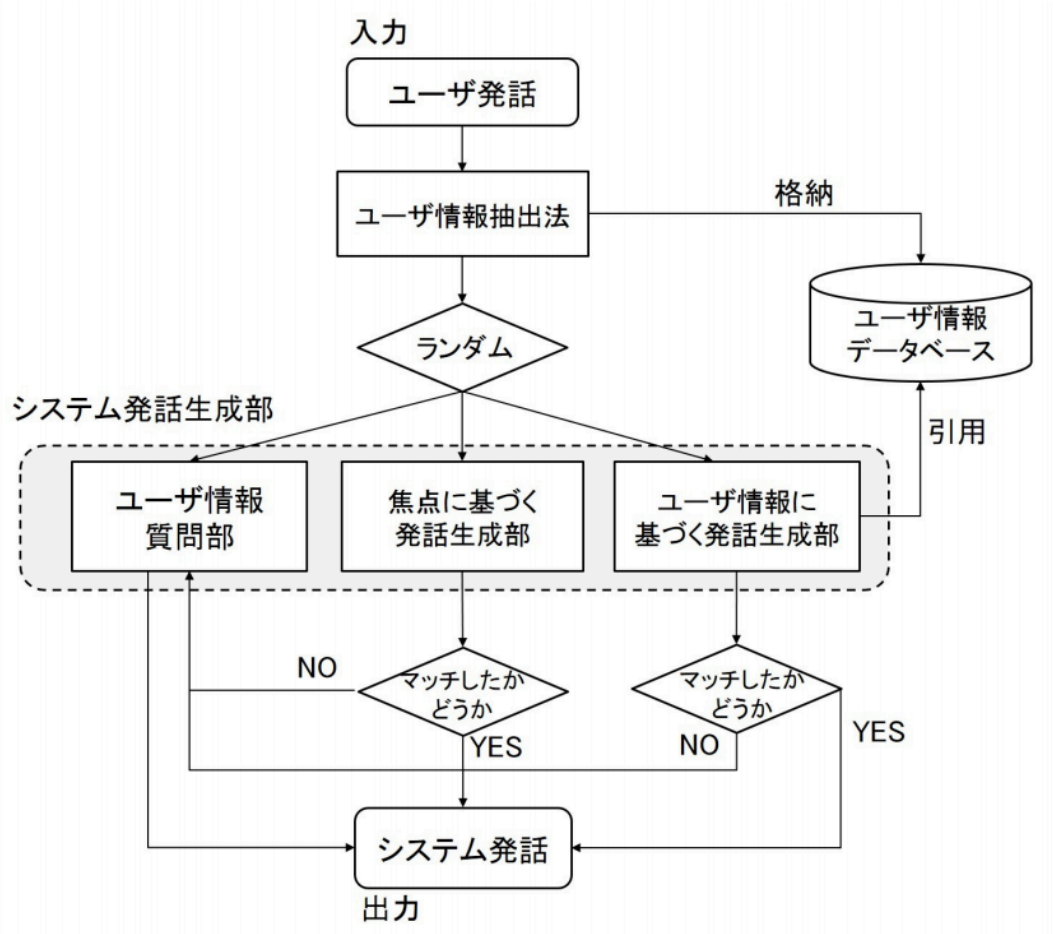
URL：https://www.jstage.jst.go.jp/article/tjsai/34/2/34_B-I65/_pdf/-char/ja

劉超然; 石井カルロス; 石黒浩. 言語・韻律情報及び対話履歴を用いた LSTM ベースのターンテイキング推定. 人工知能学会論文誌, 2019, 34.2: C-I65_1-9.
課題：turn-taking 的多 class 分類
　　Class：turn-keeping，turn-giving，question，backchannels
結果：把說話人的 ID 和發言履歷也一起 train 時，分類結果最好
　　可能是 data 小的原因，RNN 沒有效果好，LSTM 的效果最好·
技術：Word2Vec，Fast-Text，GloVe
URL：https://www.jstage.jst.go.jp/article/tjsai/34/2/34_C-I65/_pdf/-char/ja

梅澤舞菜; 入部百合絵; 北岡教英. 方言を考慮した音声言語情報に基づく高齢者認知症傾向の検出. 第 81 回全国大会講演論文集, 2019, 2019.1: 463-464.
課題：考慮多種方言的人，檢出癡呆症患者
結果：排除方言聲音的特徵檢出癡呆症的準確率跟沒有排除的比起來 10%以上好
技術：TTR
URL：file:///Users/natsuki/Downloads/IPSJ-Z81-4ZE-07.pdf

「多言語　音声」
「対話システム」

TSUNOMORI, Yuiko, et al. An Evaluation of a Chat-oriented Dialogue System that Remembers and Uses User Information over Multiple Days ユーザ情報を記憶する雑談対話システムの構築とその複数日にまたがる評価. 2020.
課題：把從對話中抽出的 user data 有效得活用在閒聊對話系統
結果：系統跨日子也記得 user information 會有好的效果
技術：CRF
Note：判斷是否在對話內容當中有 user information 的基準
　　有積極公開自己 information 的對話行為
　　有共鳴、同意、並且前面是 system 問問題的對話行為
Model：主要是使用 text 的 classification



URL：https://www.jstage.jst.go.jp/article/tjsai/35/1/35_DSI-B/_pdf/-char/ja

蔦侑磨, et al. 疑似応答を用いた雑談対話システムの自動評価. 研究報告自然言語処理 (NL), 2019, 2019.13: 1-6.

課題：在閒聊系統當中可用 BLEU 評價的自動模擬答案創造與評價結果

結果：有一些自動創造的答案與人做的答案的相關性不低，但目前來看用 BLEU 評價的話可使用的樣子。

技術：BLEU，FFNN

Note：

Model：

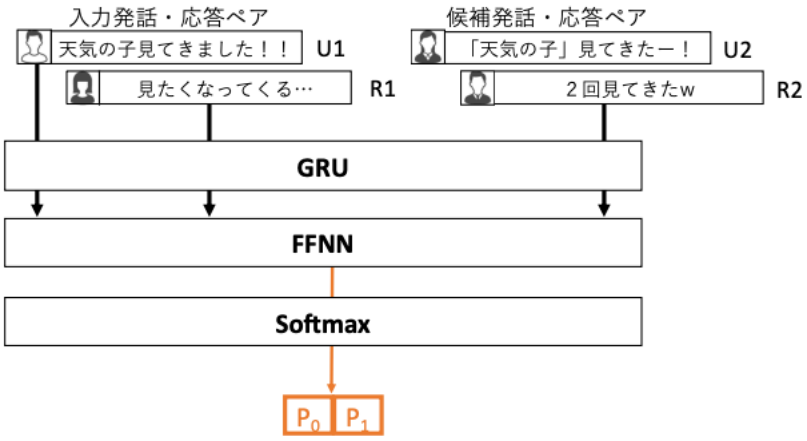


図 1　疑似応答の妥当性判定を行う分類器

Fig. 1　A classifier for judging the appropriateness of pseudo responses.

URL：http://www.tkl.iis.u-tokyo.ac.jp/new/uploads/publication_file/file/919/013984-file1.pdf

「語用論　音声」
「語用論　音声認識」
「語用　音声認識」
「語用論　対話　自然言語処理」

太田博三, et al. ポライトネス及び配慮表現コーパス作成と分析手法の一考察~ 対人関係を考慮した対話システムの適用に向けて~. SIG-SLUD, 2019, 5.01: 19-20.
報告內容：最近，在機器翻譯、對話系統等領域中，politeness 的重要性提高，較容易 annotating、較容易 customizing data 的 corpus 的需求越來越大。考慮 politeness 的 corpus 的製作是個國家專案之一，現在研究開發中。
URL：https://pj.ninjal.ac.jp/corpus_center/lrw/LRW2019_O-3-1.pdf

宮崎千明; 佐藤理史. 発話テキストへのキャラクタ性付与のための音変化表現の分類. 自然言語処理, 2019, 26.2: 407-440.
課題：音韻變化是否說話人的 character 之一
結果：137 種的音韻變化分類會 cover100 分之 80 的角色語音變化。Rule base 的這個方法造成不自然的變化了，還有改善的空間
技術：rule base
URL: https://www.jstage.jst.go.jp/article/jnlp/26/2/26_407/_pdf/-char/ja

「機械翻訳　音声」
「機械翻訳　音声認識」
「機械翻訳　スタイル」

梶原智之, et al. スタイル変換のための折り返し翻訳に基づく事前訓練. 研究報告自然言語処理 (NL), 2019, 2019.16: 1-8.

課題：用為了 style 變換的變換器訓練

結果：加上 RT 的 data 再次訓練出來，結果 10 points 以上變好，這個表示這樣的方法對轉移學習可能會有價值。

技術：Roundtrip Translation

Note : Only dealing with informal->formal. Because informal style has a lot of expression and it is more difficult to change styles with small data set.

URL: http://cl.sd.tmu.ac.jp/~kajiwara/publications/nl241-kajiwara.pdf

小田登志子. 機械翻訳と共存する外国語学習活動とは. 2019.

課題：機器翻譯和外國語習得的共存

內容： 為了 master 機器翻譯的使用，需要基本外國語知識

因為機器翻譯還無法好好考慮脈絡，所以人要重視對機器翻譯來說難的部分

複句的處理

文化方面的知識

Communication approach 等等

URL:
https://repository.tku.ac.jp/dspace/bitstream/11150/11398/1/jinbun145-03.pdf

「"machine translation" "cross culture"」

LATIF, Siddique, et al. Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends. *arXiv preprint arXiv:2001.00378*, 2020.

內容：圍繞著 Speech Recognition, Speaker Recognition, Emotion Rcognition, 整理最近流行的概念和技術。

URL: https://arxiv.org/pdf/2001.00378.pdf

《What I studied》

活性化関数

Logistic 函數

Odds:發生事情 A 和不發生事情的比率 P

P=p/1-p

Odds 的對數函數

P'=log p/1-p

把 Odds 當成一種 function

F(p)=log p/1-p

這個 function 的反函數

$$g(x) = \frac{1}{1 + e^{-x}}$$

Softmax 函數

3 次元時，x0=0，logistic 函數

$$y_2 = \frac{1}{1 + e^{-x_2}}$$

可以理解成，某一個參數跟其他參數比起來比較明顯的話，結果會靠近 1

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

hyperbolic tangent 函數

softmax 是 0<output<1

hyperbolic tangent 是 -1<output<1

實作上 hyperbolic tangent 的 output 包含 0，所以比較方便

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Note：LSTM 內部採用 tanh

Reference: https://qiita.com/Ugo-Nama/items/04814a13c9ea84978a4c#%E6%B4%BB%E6%80%A7%E5%8C%96%E9%96%A2%E6%95%B0%E3%81%82%E3%82%8C%E3%81%93%E3%82%8C

誤差逆伝播法

用一下的規則算出想求的坡度的函數（叫做 loss function）

Loss Function

E=E(y1,...,yK)

K is the number of the output units

yk is the kth output

Objective Function

$$E_{total} = \frac{1}{N}\sum_{n=1}^{N} E^n$$

Chain rule

$$\frac{\partial f}{\partial w} = \sum_{k=1}^{K} \frac{\partial f}{\partial y_k}\frac{\partial y_k}{\partial w}$$

Note：類似 cross entropy 的算法

Note:像課堂上的 backword algorithm

Reference: https://qiita.com/Ugo-Nama/items/04814a13c9ea84978a4c#%E6%B4%BB%E6%80%A7%E5%8C%96%E9%96%A2%E6%95%B0%E3%81%82%E3%82%8C%E3%81%93%E3%82%8C

Word2Vec

Note：用從 input 猜周遭的詞的虛擬 task 的方式來計算，高效率。多多少少可以追結果出來的過程。

Fast-Text

Word2vec 沒有，Fast-Text 有的功能就是 subword devide。如果詞 A 構成 B 的一部分的話，相似度會搞一點。因此可以比較好的 deal with 縮寫的詞彙。

GloVe

Word2vec 和 SVD 的特點都拿好了。

學習比較快->大的 data set OK

小的 data set 也 OK

GRU

LSTM 的 variation 之一，差別在於沒有像 LSTM 的 output gate

用有一些比較小的 data set 的時候，效果會好一點。

TTR（type of token ratio）

Unique 的 token 的總數/全 token 數

BLEU

自動評價標準之一。如果 Machine Translation System 翻譯的跟 proffesional 的人翻譯的句子是一致的話最好，這樣的假設之下，測的這些兩個的類似度的尺度就是 BLUE。

BLEU＝對短的 MT 翻譯的 penalty * n-gram 的 Geometric mean

Reference: http://www2.nict.go.jp/astrec-att/member/mutiyama/corpmt/4.pdf

《What I think & How I feel》

1. 機器翻譯的 model 都是用文字來處理，沒找到直接用語音來處理
    a. 應該是只用語言的話，比較難處理意思
    b. 但是如果有更多更多語音 data 的話，語音 data 也可以做成 word2vec，同時也可以計算 emotional feature。
    c. 也如果有說話人的 user information 的話，方言的訓練

2. 沒有看到很多重視語用論相關的機器翻譯的論文，也沒有找到直接用語音來處理
    a. 應該是因為如果要看文章或對話脈絡的話，自然語言理解、自然語言生成還是有點困難。沒有到研究語用論的地步。
    b. style 是看到好幾個研究。但看到的成功的研究主要是 rule based。
    c. politeness 相關的，還是換成口語是有點有難度的樣子。（換成 formal 的是比較有結果）
    d. 但是如果對話中的 style，politeness 的話，直接用語音或是用語音可能對結果有幫助。

3. 雖然有機器說的話對人的感覺的印象的研究，但只能用問卷調查，沒用其他科學證明
    a. 聲音高低、停頓、communication pattern（approach）等等對人影響以及印象是可能可以用別的方法來調查。
        i. 比如電波，血壓，心拍數等，可以用生理性的 data
    b. 可能心理學領域方面，別人說話的方式、口氣、內容、個性等等的要素對聽話者的影響。調整自然語言生成 model 時，可能可以加其他領域的 data 過來的 feature。

4. 論文裡常出現的技術沒有差很多。
    a. 看到了的是改一點在課堂上學到的技術、上述 what I studied 裡面寫的技術的
    b. 如果把基礎打好，研究論文沒有那麼困難。越熟悉常用的技術，越順利的看完論文，越容易看懂論文的內容。

5. 自然語言處理、自然語言生成、語言辨識、語言生成的領域裡面，沒有找到文化相關的研究。
    a. 可能語言的表面上，沒有看得到的文化
    b. 但常用的單詞排名代表母語者的思考方式的一部分，把單詞統計下來，跟麒泰語言比較，如果有不一樣的地方的話，可能思考方式（習慣，思想，文化）不一樣。
    c. 關於跨文化 communication approach、有一些地域的人常會用諷刺的方式來指責別人，另有一些地域的人常會直接指責別人，再另有一些地域的人會先誇

獎幾個地方再指責等，每個地域的 communication approach 不一樣。這樣的 communication approach 是每個語言的對話比較起來可能可以找出各個語言的 communication approach 的特徵（習慣，思想，文化）。

d. 關於跨文化的 politeness 相關的 communication approach，人的關係、口氣、情緒、場合、說話內容等，人要決定怎麼講的時候很多因素會有關係而且每個對話重視的因素可能不一樣。比如工作場合並且工作內容的話，人的關係應該是決定說話方式的最重要的原因，說話方式可能是很 polite。但工作場合並且自己做錯的內容的話，可能口氣的變化比較多，而且可能在對話裡的有一些不分很 impolite（因為可能會再現被罵的情景等）。這樣的話影響到的因素，被影響到的因素跟上個例子不一樣。
e. 如果語音的話，只處理文字的時候，抓不到的 feature 可以抓到。用語音可能找得到至今自然語言處理沒辦法處理的事情。

《What is my idea》

1. 用語言的方言 data 來加強方言特徵，提高方言文字生成語音生成的準確度
    a. 可能可以實驗每一個方言中的差異、類似度會不會有效。
    b. 可能可以統計方言之間的音素的差異和重疊
2. 用 user information 的 data 來，提高檢出 emotion 的準確度
    a. 可能可以訓練 user 平常的特徵和有 emotion 時的特徵
    b. 可能可以用 OOV 的技術應用在這 task 上
3. 用 user information 的 data 來，提高方言的語言理解的準確度的準確度
    a. 出身地，長大的地方，工作的地方等的資訊可能對它很有效
4. 用 user information 的 data 來，提高方言的文字生成語音生成的準確度的準確度
    a. 出身地，長大的地方，工作的地方等的資訊可能對它很有效
5. 用語音檢出的 Emotion 幫助 metaphor 的 disambiguation
    a. 用 metaphor 的時候，可能有一些 feature
    b. 如果找到這個 feature，對 metaphor 的 disambiguation 有幫助
6. 用語音檢出的 Emotion 幫助 style 變換，politeness 調整
    a. Emotion 特徵可能有一些 feature 對它們有幫助的
    b. 可能語音簡出的 emotion 特徵 Style，politeness 程度上的變換
    c. 加上另外收集的 style，politeness 的程度上的 data，可能效果更好
7. 腦波、血壓、心拍數等的 data 可能可以跟語音 data 一起紀錄，一起處理是否各種 task 的效果會不會好。
    a. Emotion detection
    b. Speech personalizing
    c. 怎麼樣人可以比較舒服的跟機器溝通的問題
    d. 機器翻譯給人的印象改善
    e. Personalizing speech generation
    f. Personalizing language generation
    g. Personalizing style transformation
    h. Personalizing politeness control
    i. 等等
8. 用各個語言的語音的對話 data 來比較，communication approach 的各國的 feature（差別和類似的點）
    a. 可能需要跟 situation 綁定
    b. 可能需要跟人的關係綁定
    c. 可能需要跟口氣綁定
    d. 可能需要跟 emotion 綁定
    e. 可能需要跟 topic 綁定
    f. 可能需要跟口語常用 idiom 綁定
    g. 可能需要跟口語常用單詞綁定

h. 可能需要跟口語常用 metaphor 綁定
i. 可能需要跟口語說法綁定
j. 可能需要在那個地域大眾擁有的思想和話題的 data 綁定
k. 可能需要其他領域（社會科學，腦科學，心理學）的統計 data 拿來用