Lingxi: A Diversity-aware Chinese Modern Poetry Generation System

摘要

好的中文現代詩，遵循陌生化原則，低頻詞和模糊的句子被認為是新穎有創意的。本論文提出Lingxi希望可以生出這種詩。

Nucleus Sampling with ramdomized heads: 將預測分布的高頻"頭"隨機化，強調低頻詞，增加新穎性，調整高頻率波參數，可以控制分布變化，實現所謂Diversity-aware，詞彙過濾有隨機性(近80%)，但還是可以生成流暢詩歌，且符合新穎性目的。

語意相似拒採樣: 取符合條件的樣本(拒採樣, Rejection Sampling)，創造互多信息輸入給模型，和上次說明的KEE目的相似。

簡介

Nucleus Sampling(Top-p Sampling): 使用隨機抽樣(stochastic sampling)取代beam search緩解文本退化(text degeneration)，截斷低頻保證品質，in PGT(Poem Generation Task) 扔會生成無趣且重複的詩，沒解決新詩的陌生化原則。

此論文之場景:指輸入標題OR KEY Word，生成完整詩歌段落

Pre-train LM

發布一名為GPT-LyricCN的Pre-train LM:

- pre-train data: 3500本已出版中文小說
- fine-tuning data: 22萬首現代詩和中文音樂歌詞
- 訓練方法: 同GPT-2

詞表取得:

1. THULAC分詞處理語料，詞頻排序取Top-90%為basic vocab(共12643)
2. OOV以basic vocab分詞，如無法處理，將該詞加入basic vocab(共23296)，如有多種切法，取maximum likehood product
3. 再次對basic vocab排序取Top-n為final vocab(共17589)

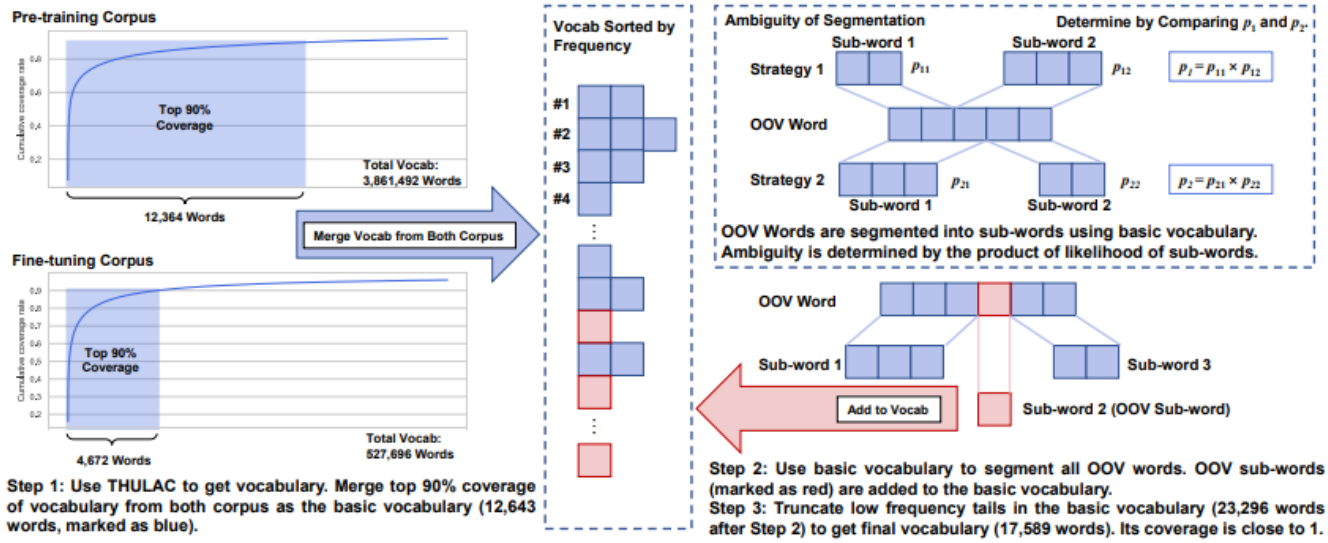


Figure 4: Illustration of the corpus preprocessing.

Diversity-aware Sampling(多樣性抽樣)

Controllable Diversity by Permutating the “Head” of Predicted Distribution(通過置換預測分佈的“頭部(高頻)”來控制多樣性)

警告：以下說明以斜體表示內容有可能錯誤，有其他解釋空間，請參考原論文或跟我討論你的看法

Top-p Sampling ,Top-k Sampling: 克服文本退化，better than beam-search，但在PGT無效。

解法: 先取一次Top-p Sampling(q)，取head的過閥值的部分(最大機率/n)然後重新分配機率密度，*head外的在做一次Top-p Sampling(q)取head(可能錯誤:有可能不只是head外，head內也可能會)*，在這個新的分布隨機抽樣。

Semantic-similarity-based Rejection Sampling Algorithm(基於語義相似性的拒絕採樣算法)

使用機率低的詞Decode實際會出現主題偏離的情形。

解法:

1. 對前M個token進行N次取樣。
2. 對每個樣本token計算和輸入主題(標題、關鍵字)的相似度(BERT Sentence embedding 算餘弦相似度(夾角))，接受最大值樣本token。
3. 將此token送model，decode出剩餘token。

Demonstration and Evaluation

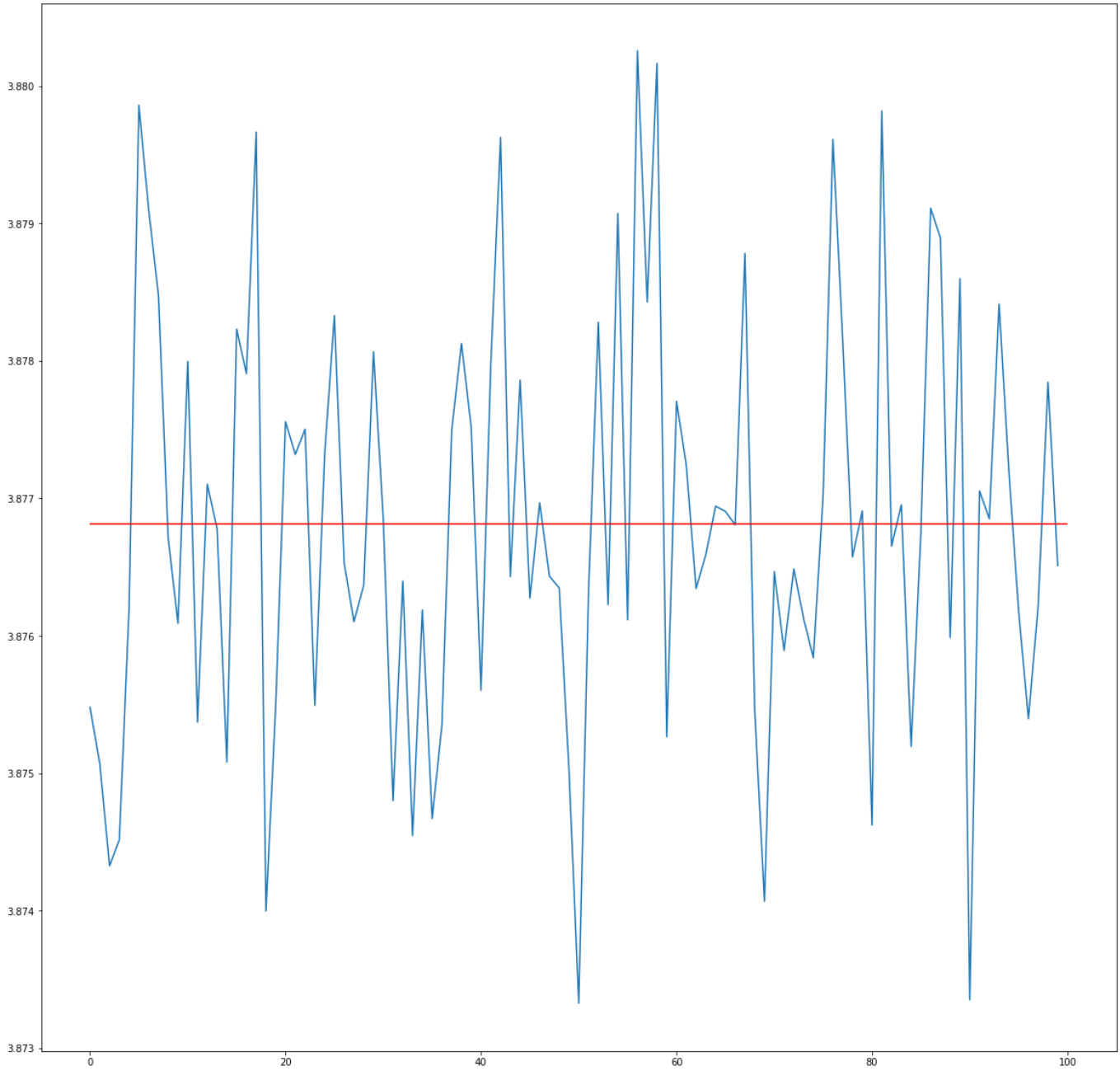
evaluation paradigm

PPL(Perplexity, 混淆度): 生成樣本的總體流暢度（分數越低表示流暢度越高，但越無聊）

Self-BLEU: 不同樣本之間的多樣性(分數越低表示多樣性越高)

Zipf: 反映詞頻分佈特徵(分數越低表示詞頻分佈越平坦，多樣性越高)

Repetition entropy: 分數越高表示重複越少，多樣性越高，$\mathbb{E} \left { - \log_{}{p\left (x \right) } \right } $，p為生成樣本中的詞頻分佈



Rhyming entropy: 分數越高表示多樣性越高，但押韻越少，$\mathbb{E} \left { - \log_{}{p_{rhyme}\left (x \right) } \right }$，p為生成樣本中的押韻頻率分佈

line length: 生成樣本中每行詩出現的單詞平均數，越高分代表這首詩越長，且使用更豐富的詞彙

Method	PPL	Self-BLEU 4	Self-BLEU 5	Zipf Coef.	Rep. Entropy	Rhy. Entropy	Line Length
Human	16.75	0.45	0.32	0.90	3.68	1.50	7.42
NS, $p = 0.70$	2.70	0.65	0.51	**0.87**	3.14	**1.52**	5.84
NS, $p = 0.90$	6.80	0.51	0.36	0.82	3.59	1.85	6.38
Top-k, $k = 200$	7.31	**0.50**	**0.35**	0.81	**3.61**	1.86	6.66
Pure sampling ($p = 1.00$)	**18.83**	**0.40**	0.25	0.80	3.87	2.03	**7.05**
IQR-IP, $\rho = 10.00$	6.60	0.55	0.38	**0.81**	**3.70**	**1.85**	**7.24**
IQR-IP, $\rho = 5.00$	9.54	0.51	0.34	0.76	3.86	1.88	7.90
IQR-IP, $\rho = 3.00$	12.13	0.48	**0.32**	0.73	3.96	1.91	8.29
IQR-IP, $\rho = 1.50$	**18.14**	**0.43**	0.27	0.70	4.08	1.92	9.33
NS-RH, $p = 0.80, q = 0.20$	4.00	0.58	0.43	**0.84**	3.37	**1.70**	6.09
NS-RH, $p = 0.80, q = 0.40$	5.06	0.56	0.40	0.83	3.52	1.79	6.28
NS-RH, $p = 0.80, q = 0.60$	8.82	**0.50**	**0.34**	0.78	**3.80**	1.95	**6.69**
NS-RH, $p = 0.80, q = 0.80$	**28.71**	0.35	0.21	0.70	4.17	2.00	8.67
NS-RH, $p = 0.90, q = 0.20$	6.97	0.51	0.36	**0.82**	3.60	**1.87**	6.40
NS-RH, $p = 0.90, q = 0.40$	8.80	0.49	**0.33**	0.81	**3.72**	1.93	6.51
NS-RH, $p = 0.90, q = 0.60$	**14.45**	**0.44**	0.27	0.75	3.95	2.02	**7.08**
NS-RH, $p = 0.90, q = 0.80$	41.54	0.30	0.17	0.68	4.24	2.01	9.25

Table 1: Statistical evaluation for selected decoding parameters (*metric closer to human text is better and in bold*). Lower self-BLEU score, lower Zipf coefficient and higher repetition entropy indicates *higher diversity*.

evaluation

PPL: 傳統隨機採樣方法生成的樣本嚴重退化，因為它們的PPL（NS，p = 0.70 或 0.90 和 top-k，k = 200）遠低於人類，更高的PPL可能被認為是詩意的，當 p = 1.00 時才能達到接近人類指標的最大 PPL，在此狀態下模型會無條件考慮低機率詞，破壞詩的結構。相比之下，NS-RH的PPL到達人類水平，有條件考慮低機率詞，不會破壞詩的結構。輸出Robustness可透過p、q調整。NS-RH 在 p = 0.90, q = 0.80 並且生成了滿意的樣本，代表其Top80%的選擇是隨機的，扔然能生成流暢詩歌，在 p = 0.80, q = 0.80(砍掉"尾")更近似人類。

Self-BLEU: 實現人類指標

Zipf: 達到更低值，代表生成所使用的詞彙更加多樣化

Repetition entropy: 達到更高值，代表重複性更低

Rhyming entropy: 高於人類，表明犧牲押韻能力換取多樣性

line length: 實現與人類文本相似的詩歌格式

消融測試:拒絕採樣算法

圖表左方為使用拒絕採樣算法的結果，右方反之。可以看見使用拒絕採樣算法的模型生成前後文的BLEU分數更高，有助於保持主題的一致性。

Diversity	BLEU ↑ (×0.01)	
	w/ RJ	w/o RJ
$\rho = 10.00$	0.67	0.50
$\rho = 5.00$	0.67	0.45
$\rho = 3.00$	0.65	0.39
$\rho = 1.50$	0.62	0.34
$q = 0.20$	0.97	0.56
$q = 0.40$	0.70	0.41
$q = 0.60$	0.53	0.31
$q = 0.80$	0.37	0.23

人類評價

和其他現有的生成系統(Youling(屬性控制), XiaoIce(img2poem))評價，以"泉水"為輸入，生50首詩，每項1~5分↑。

人員組成: 研究生、研究小組顧問、音樂家和作曲家

Fluency: 流暢性，側重語法與語言學

Novelty: 新穎性，詩意和創造性

Coherence: 連貫性，詩歌與標題連貫性

Overall: 以上平均

System	Fluency	Novelty	Coherence	Overall
Youling (Zhang et al., 2020)	4.37	4.16	4.19	4.24
XiaoIce (Cheng et al., 2018)	**4.46**	4.12	4.12	4.23
Lingxi (*ours*, $q = 0.60$)	4.43	**4.31**	**4.23**	**4.32**

總結

1. Lingxi: Diversity-aware 中文現代詩生成器
2. NS-RH: 可控多樣性解碼方法 3: Semantic-similarity-based Rejection Sampling: 緩解主題偏移

將分布高頻隨機化，顯著增加生成詩歌新穎性。