A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu[1], Wai Lam[1], Anthony Man-Cho So[1], Bei Shi[2]
[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong
[2]AI Lab, Tencent

https://github.com/fuzihaofzh/repetition-problem-nlg

AAAI 2021

Outline

➡Introduction

- Theoretical Analysis
- Rebalanced Encoding
- Experiments
- Conclusions

Text Generation

- Text generation tasks aim at generating human-readable text for specific tasks.
    - e.g. Machine Translation, Summarization, Data-to-text Generation, Language Modeling, and etc.
- Frameworks
    - Encoder-Decoder
    - Language Model

Repetition Problem

- Repetition problem

Tough it is still unfinished, $\underbrace{\textbf{but I like it but I like it but I like}}_{\text{Repetition}}$ **...**

- Some conjectures of the reason:
  - The model architectures (Vig 2018; Holtzman et al. 2020)
  - The gap between sampling methods and the real human language (Holtzman et al. 2020)
  - Reliance on the fixed corpora cannot fulfill the real goal of using the language. (Choi 2018)
  - Likelihood maximizing decoding. (Welleck et al. 2020)

Our Contribution

- **New theoretical framework**
- **More understanding of our language**
  - Our language contains too many words predicting the same word (i.e. high inflow words) as the subsequent word with high probability.
  - Repetition problem is caused by the high inflow words in our language.
- **Unified understanding of existing algorithms**
  - including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A new encoding method**
  - Rebalanced encoding approach solves the high inflow problem and thus alleviates the repetition problem.

Outline

✓Introduction

➡Theoretical Analysis

- Rebalanced Encoding
- Experiments
- Conclusions

Definitions

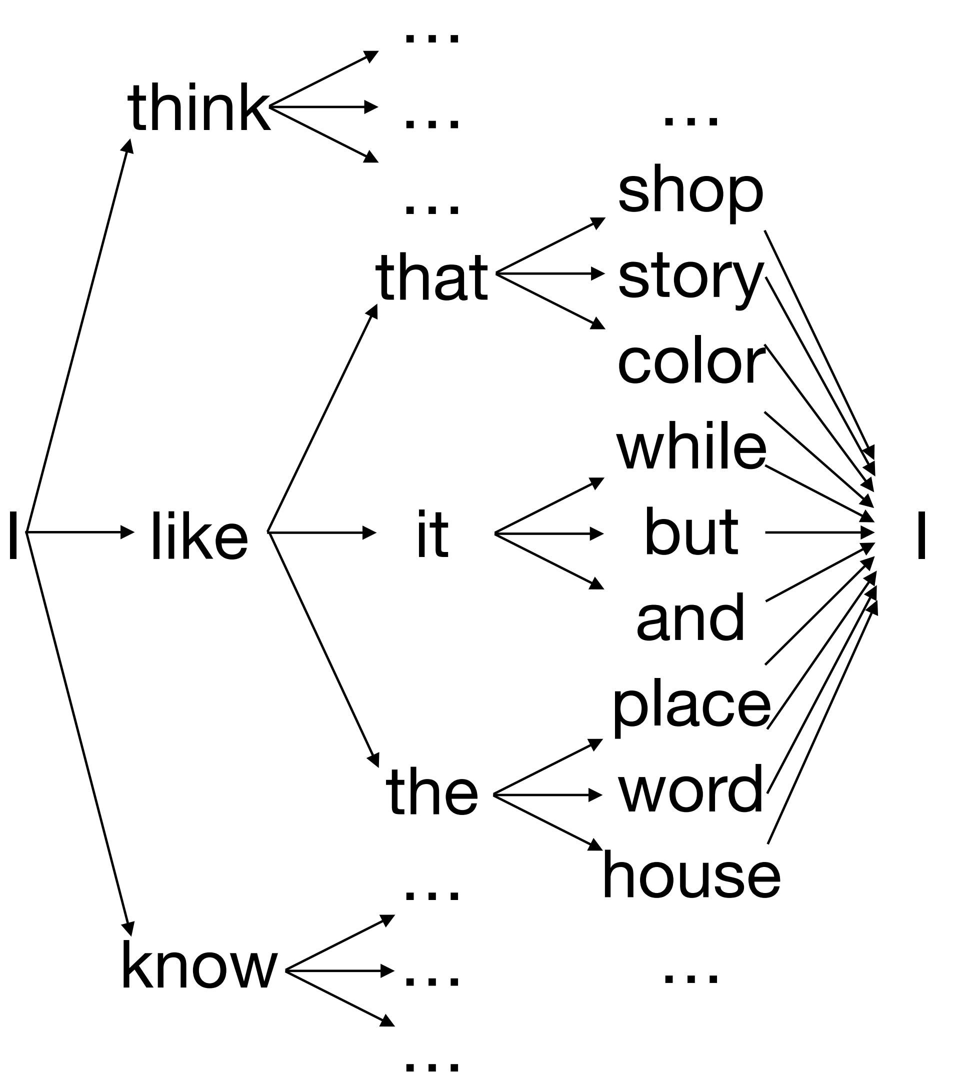
Definition 2.1 (Markov Generation Model)

- A Markov generation model predicts a word only based on the previous word, which can be denoted as $p_\xi(w_i \mid w_{i-1}) \approx p_\theta(w_i \mid w_{i-1}, \cdots, w_1, x)$, where ξ stands for parameters of the Markov generation model.
- We denote the transition matrix as $A \in \mathbb{R}^{(n+1)\times(n+1)}$, in which n is the vocabulary size. $A_{ij} \geq 0, \sum_{j=1}^{n+1} A_{ij} = 1, \forall i \in [1, n+1]$. (n+1)th word is the EOS tag. A can be written as $A = \begin{bmatrix} B, b \\ 0, 1 \end{bmatrix}$, in which $B \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^{n\times 1}$.

Definitions

Definition 2.2 (Average Repetition Probability)

- The sparsity of matrix B: $\zeta = \frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} 1(B_{ij} > 0)$
- For a word transits at the kth step, it generates $(\zeta n)^k$ paths and has $(\zeta n)^k / n$ paths that transit back to itself on average
- The probability for the ith word loops back to itself after k steps is B_{ii}^k
- The average probability for each path is $\frac{nB_{ii}^k}{(\zeta n)^k}$
- The average probability for all loops to repeat again is $(\frac{nB_{ii}^k}{(\zeta n)^k})^2 \cdot \frac{(\zeta n)^k}{n}$



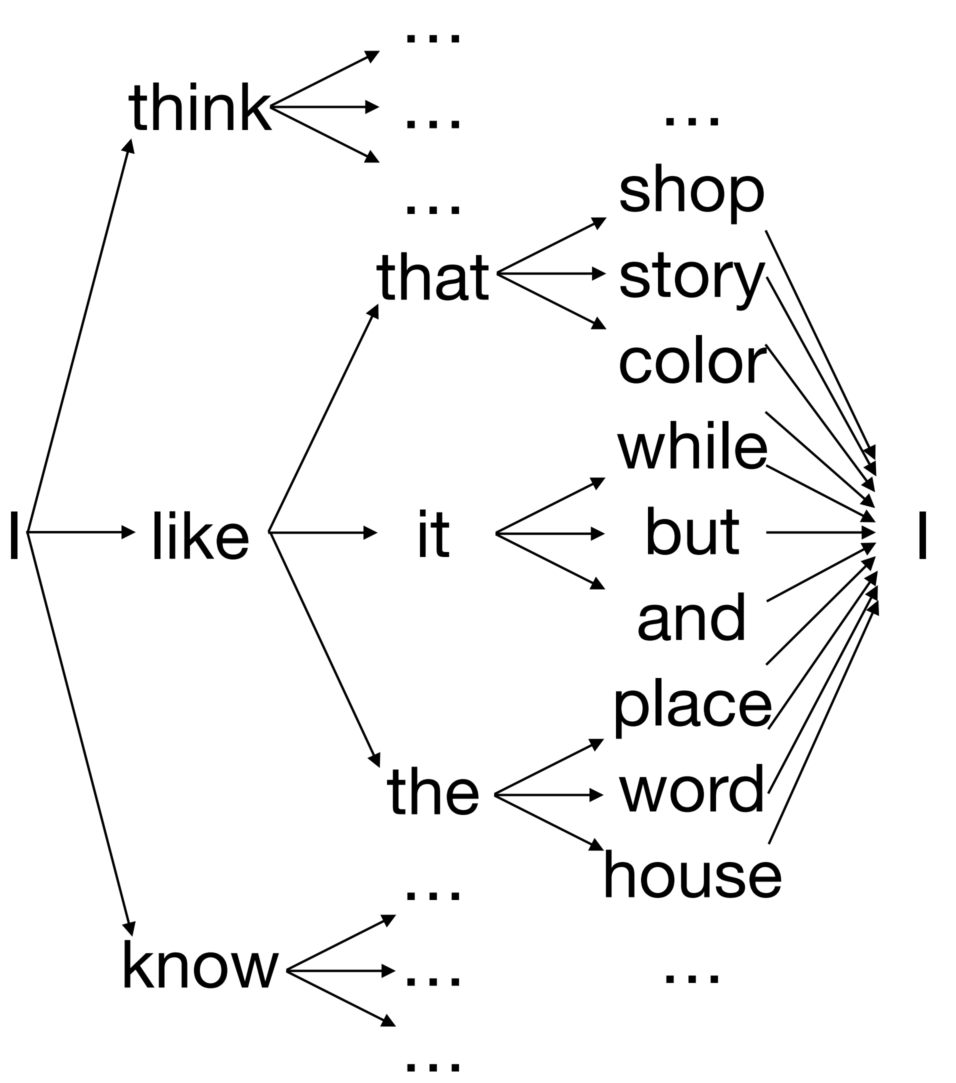
Definitions

Definition 2.2 (Average Repetition Probability)

- The average probability sum for all words in all steps is $\sum_{k=1}^{\infty} \operatorname{tr}(\frac{B^{2k}}{\zeta^k n^{k-1}})$
- The average probability for each word is $R = \frac{1}{n} \sum_{k=1}^{\infty} \operatorname{tr}(\frac{B^{2k}}{\zeta^k n^{k-1}}) = \sum_{k=1}^{\infty} \operatorname{tr}(\frac{B^{2k}}{\zeta^k n^k})$

- **(Average Repetition Probability)** Given a Markov generation model with sub-transition matrix $B \in \mathbb{R}^{n \times n}$ and non-zero probability ζ , the Average Repetition Probability (ARP) is defined as:

$$R = \sum_{k=1}^{\infty} \operatorname{tr}(\frac{B^{2k}}{\zeta^k n^k})$$



Understanding Existing Method

- General procedure for all methods:

$$w \sim T(p)$$

where p is the original word probability and T is a specific transformation.

- **Stochastic sampling** uses the identity operator, $T_s(p) = p$.
- **Greedy sampling** changes the probability of the most probable word to 1 and sets others' probabilities to 0 which can be denoted as $T_g(p)_i = 1(i = \arg\max_j p_j)$.
- **Topk sampling** sets all word probabilities to 0 except top k words and rescales it to a new distribution. It can be denoted as $T_k(p)_i = 1(i \in K)p_i / \sum_{j \in K} p_j$, in which K is the

  Topk probability word set: $K = \arg\max_K \sum_{j \in K} p_j, s.t. |K| = k$.

Understanding Existing Method

- **Nucleus sampling** sets all word probabilities to 0 except for words with probabilities sum to larger than p. It can be denoted as $T_n(p)_i = 1(i \in N)p_i / \sum_{j \in N} p_j$, in which N is the smallest word set with probability sum larger than p: $N = \arg\min_N |N|, s.t. \sum_{j \in N} p_j \geq p$.
- **Temperature sampling** rescales the probability with a temperature parameter t and can be denoted as $T_t(p)_i = \exp((\log p_i)/t) / \sum_j \exp((\log p_j)/t)$.
- **Length penalty** simply enlarge the probability of the EOS tag with a constant β which can be denoted as $T_l(p)_i = \exp(\tilde{\ell}_i) / \sum_j \exp(\tilde{\ell}_j)$, in which $\tilde{\ell}_i = \ell_i + 1(i = n + 1)\beta$ and ℓ is the logits vector calculated by the model.

ARP Upper Bound

Theorem 1. If $\zeta n > \rho(B^2)$,

$$R = \mathrm{tr}(B^2(\zeta nI - B^2)^{-1}) \leq \frac{\|B^2\|_*}{\sigma_n(\zeta nI - B^2)}$$

in which $\| \cdot \|_*$ is the nuclear norm; σ_n denotes the smallest singular value; $\rho(\,\cdot\,)$ is the spectral radius of the matrix; I is the identity matrix.

ARP Upper Bound

Corollary 1.1.

$$R \leq \frac{\sqrt{r}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(B_{ij}-\mu_i)^2+\sum_{i=1}^{n}(1-b_i)^2\right)}{\sigma_n(\zeta nI - B^2)}$$

where r is the rank of B^2 and $\mu_i = \frac{\sum_{k=1}^{n} B_{ik}}{n}$ is the mean of each row of B.

It can be concluded that the upper bound of R decreases as the variance of B_{ij} decreases.

Discussion

<u>Greedy Sampling</u>: only one word has a probability of 1 with others being 0. It has high variance.

<u>Stochastic sampling</u>: always achieves the smallest variance. However, due to the long tail effect, it cannot be used since it has a very high probability of sampling low probability words.

ARP Upper Bound

Corollary 1.1.

$$R \leq \frac{\sqrt{r}(\sum_{i=1}^{n}\sum_{j=1}^{n}(B_{ij}-\mu_i)^2+\sum_{i=1}^{n}(1-b_i)^2)}{\sigma_n(\zeta nI - B^2)}$$

where r is the rank of B^2 and $\mu_i = \frac{\sum_{k=1}^{n} B_{ik}}{n}$ is the mean of each row of B.

It can be concluded that the upper bound of R decreases as the variance of B_{ij} decreases.

Discussion

Temperature sampling: controls the variance by changing the temperature parameter to alleviates the repetition problem.

Topk/Nucleus sampling: not directly alleviate the repetition problem. It actually solves the long-tail problem of stochastic sampling or Temperature sampling. If the temperature is fixed, using Topk/Nucleus sampling can even make the repetition problem worse.

ARP Upper Bound

Corollary 1.2. If $\zeta nI - B^2$ is a diagonally dominant matrix,

$$R \leq \frac{\|B^2\|_*}{\min_{1\leq i\leq n}\{\frac{1}{2}(\zeta n - \underbrace{\sum_{j=1}^{n}(B^2)_{ij}}_{outflow}) + \frac{1}{2}(\zeta n - \underbrace{\sum_{k=1}^{n}(B^2)_{ki}}_{inflow})\}}$$

<u>inflow</u>: the probability sum of all words that take a word as the subsequent word. If it is too big, the upper bound can be magnified extensively and even fails to limit the ARP.

This Corollary theoretically justifies the claim that high inflow words causes the repetition problem.

Extend to General Models

(Average Repetition Probability for General Generation Model) For a general generation model, the sub-transition matrix for the kth step can be expressed as $B'_k = B + T_k$, in which $T_k \in \mathbb{R}^{n \times n}$ is a perturbation matrix and each element for $T_{k_{(ij)}}$ is independently distributed with mean 0 and we assume that the variance is controlled as $\delta^2 < \frac{1}{n}$. The general average repetition probability is defined as:

$$R' = \sum_{r=1}^{\infty} \mathrm{tr}(\frac{\prod_{k=1}^{2r} B'_k}{\zeta^r n^r})$$

Extend to General Models

Theorem 2. For a general generation model, if $\sum_{i=1}^{n} B_{ij}^2 < 1, \zeta n > 4$, then for every constant $a > 0$ we have:

$$\Pr(|R - R'| \geq a) \leq \frac{3\zeta n \delta^2}{a^2(\zeta n - 4)(\zeta n - 1)}$$

Given a generation model that has a different transition matrix B_k' at each step, if the transition matrix does not deviate a lot from that in the Markov generation model, the deviation of ARP is bounded with high probability.

Outline

✓Introduction

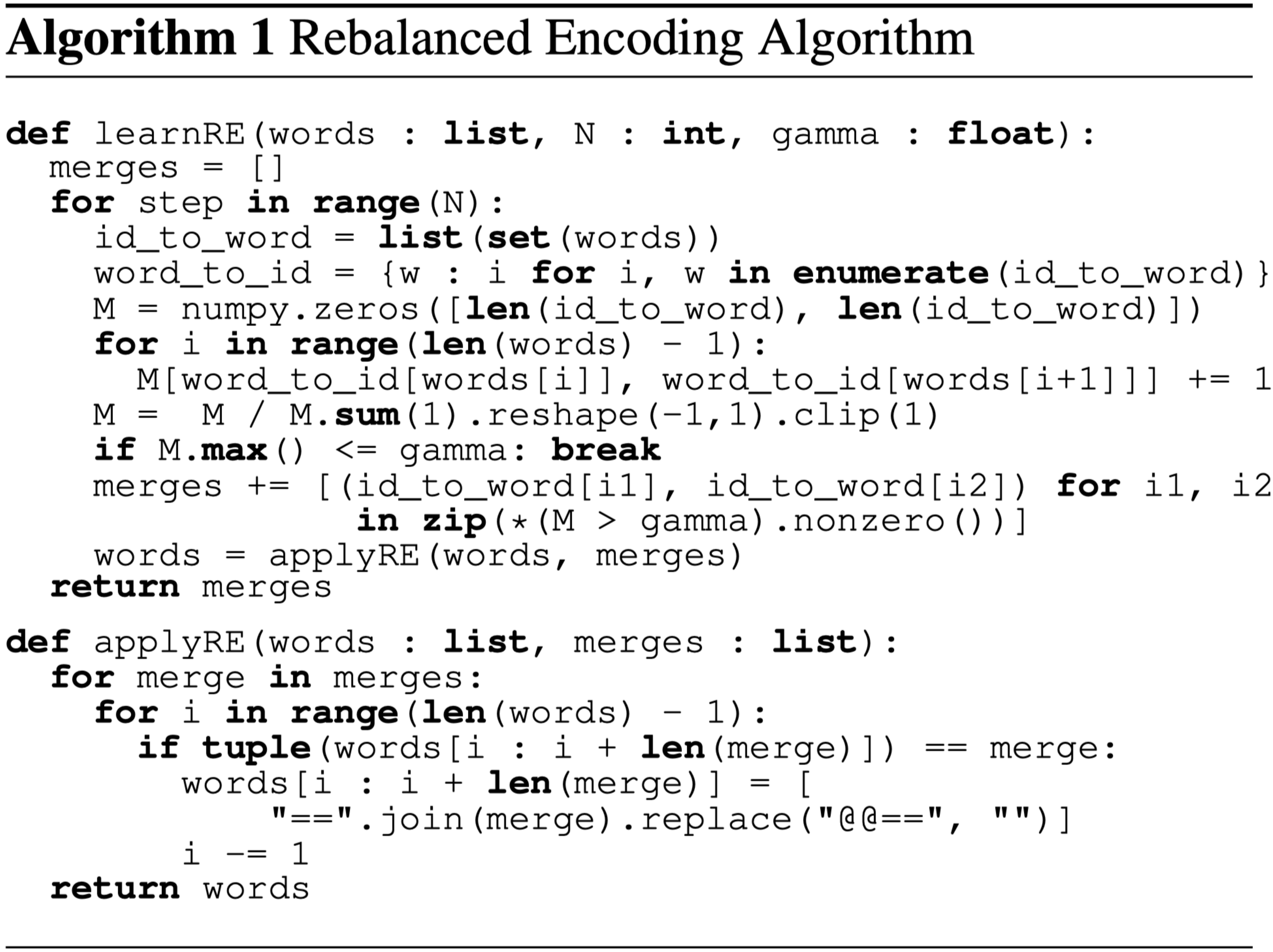
✓Theoretical Analysis

➡Rebalanced Encoding

- Experiments
- Conclusions

Rebalanced Encoding

- From Corollary 1.2, we know that the high inflow words lead to the repetition problem.
- The high inflow words is the natural of our language.
- Solution: change the word encoding.



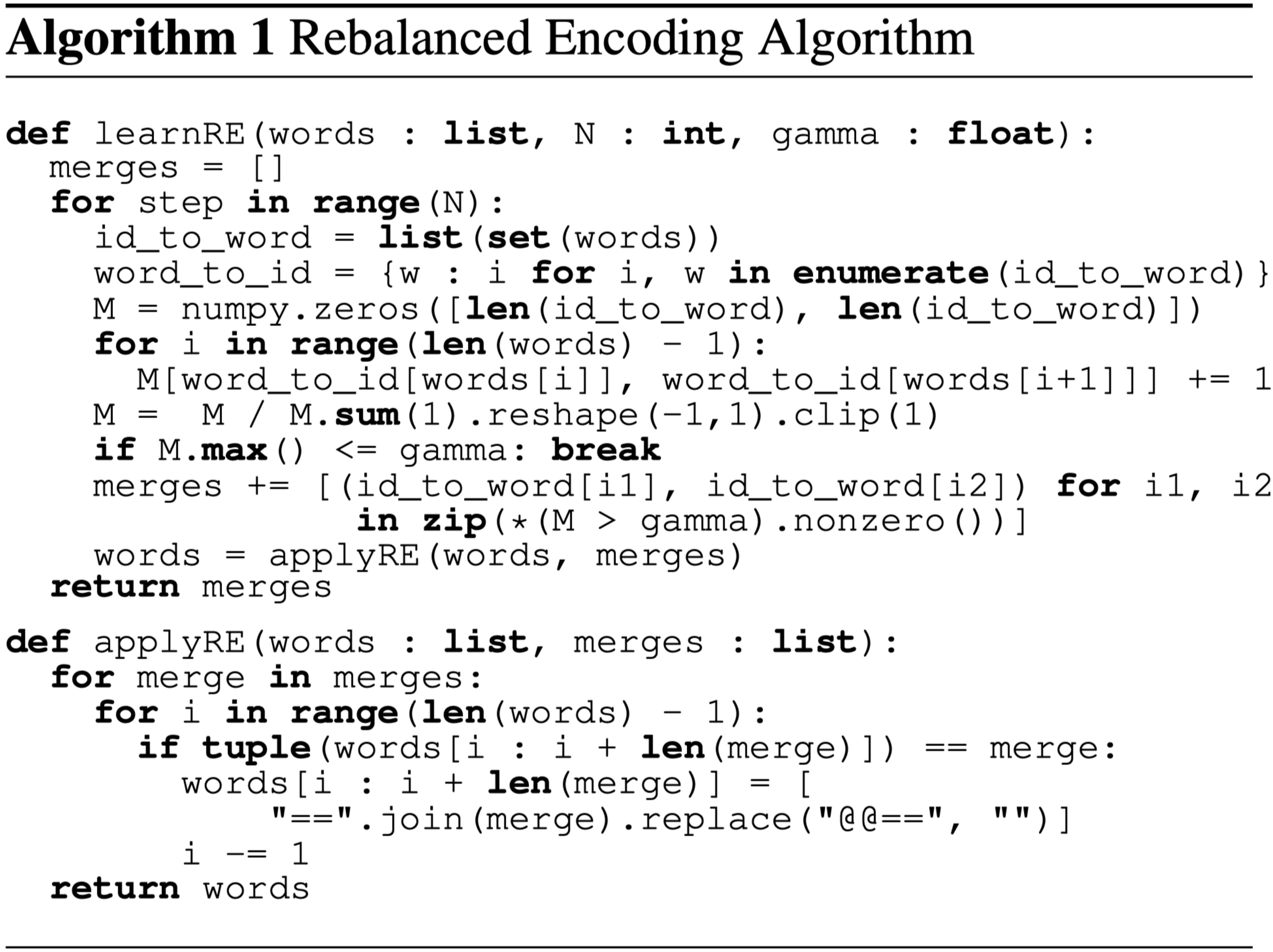
Algorithm 1 Rebalanced Encoding Algorithm

```
def learnRE(words : list, N : int, gamma : float):
  merges = []
  for step in range(N):
    id_to_word = list(set(words))
    word_to_id = {w : i for i, w in enumerate(id_to_word)}
    M = numpy.zeros([len(id_to_word), len(id_to_word)])
    for i in range(len(words) - 1):
      M[word_to_id[words[i]], word_to_id[words[i+1]]] += 1
    M =  M / M.sum(1).reshape(-1,1).clip(1)
    if M.max() <= gamma: break
    merges += [(id_to_word[i1], id_to_word[i2]) for i1, i2
                in zip(*(M > gamma).nonzero())]
    words = applyRE(words, merges)
  return merges

def applyRE(words : list, merges : list):
  for merge in merges:
    for i in range(len(words) - 1):
      if tuple(words[i : i + len(merge)]) == merge:
        words[i : i + len(merge)] = [
            "==".join(merge).replace("@@==", "")]
        i -= 1
  return words
```

Rebalanced Encoding

- makes a statistical transition matrix with the encoded training text.
- It picks high inflow pairs that have transition probability higher than a threshold γ and merges the word pair as a whole word.
  - If two words are split from BPE, we simply merge them and remove the BPE tags. E.g. for the word pair ``(de@@, crease)", we replace all ``de@@ crease" to ``decrease".
  - If the two words are words that have not been split by BPE, we merge them by adding a ``==" tag. E.g. for the word pair ``(involved, in)", we replace all ``involved in" to ``involved==in".
- We rebuild the transition matrix and repeat the above procedure until all the probabilities in the transition matrix are less than γ or we reach a specific iteration epoch.



Algorithm 1 Rebalanced Encoding Algorithm

```
def learnRE(words : list, N : int, gamma : float):
  merges = []
  for step in range(N):
    id_to_word = list(set(words))
    word_to_id = {w : i for i, w in enumerate(id_to_word)}
    M = numpy.zeros([len(id_to_word), len(id_to_word)])
    for i in range(len(words) - 1):
      M[word_to_id[words[i]], word_to_id[words[i+1]]] += 1
    M =  M / M.sum(1).reshape(-1,1).clip(1)
    if M.max() <= gamma: break
    merges += [(id_to_word[i1], id_to_word[i2]) for i1, i2
                  in zip(*(M > gamma).nonzero())]
    words = applyRE(words, merges)
  return merges

def applyRE(words : list, merges : list):
  for merge in merges:
    for i in range(len(words) - 1):
      if tuple(words[i : i + len(merge)]) == merge:
        words[i : i + len(merge)] = [
            "==".join(merge).replace("@@==", "")]
        i -= 1
  return words
```

Outline

✓Introduction

✓Theoretical Analysis

✓Rebalanced Encoding

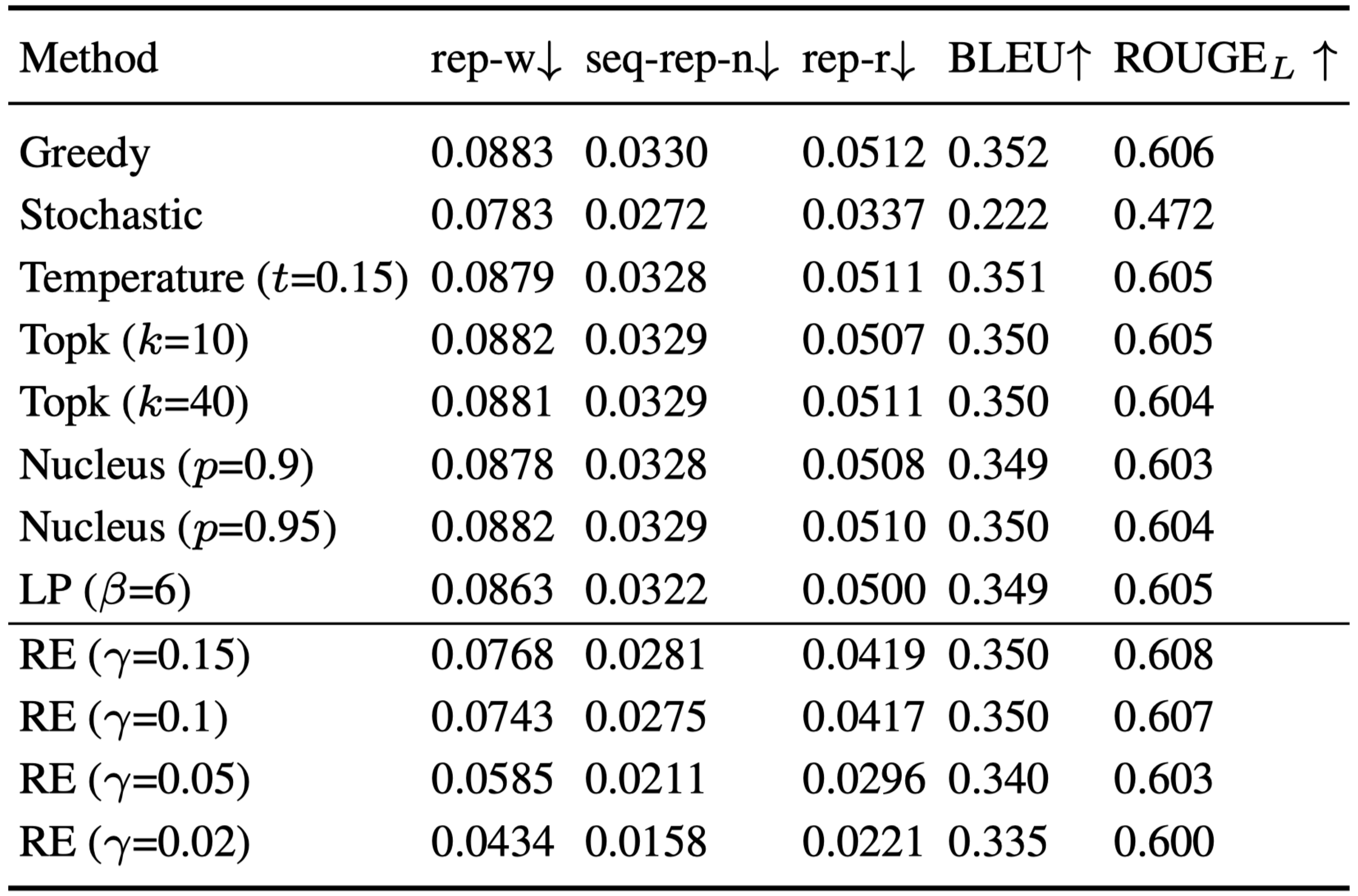
➡Experiments

- Conclusions

Experiments - Main

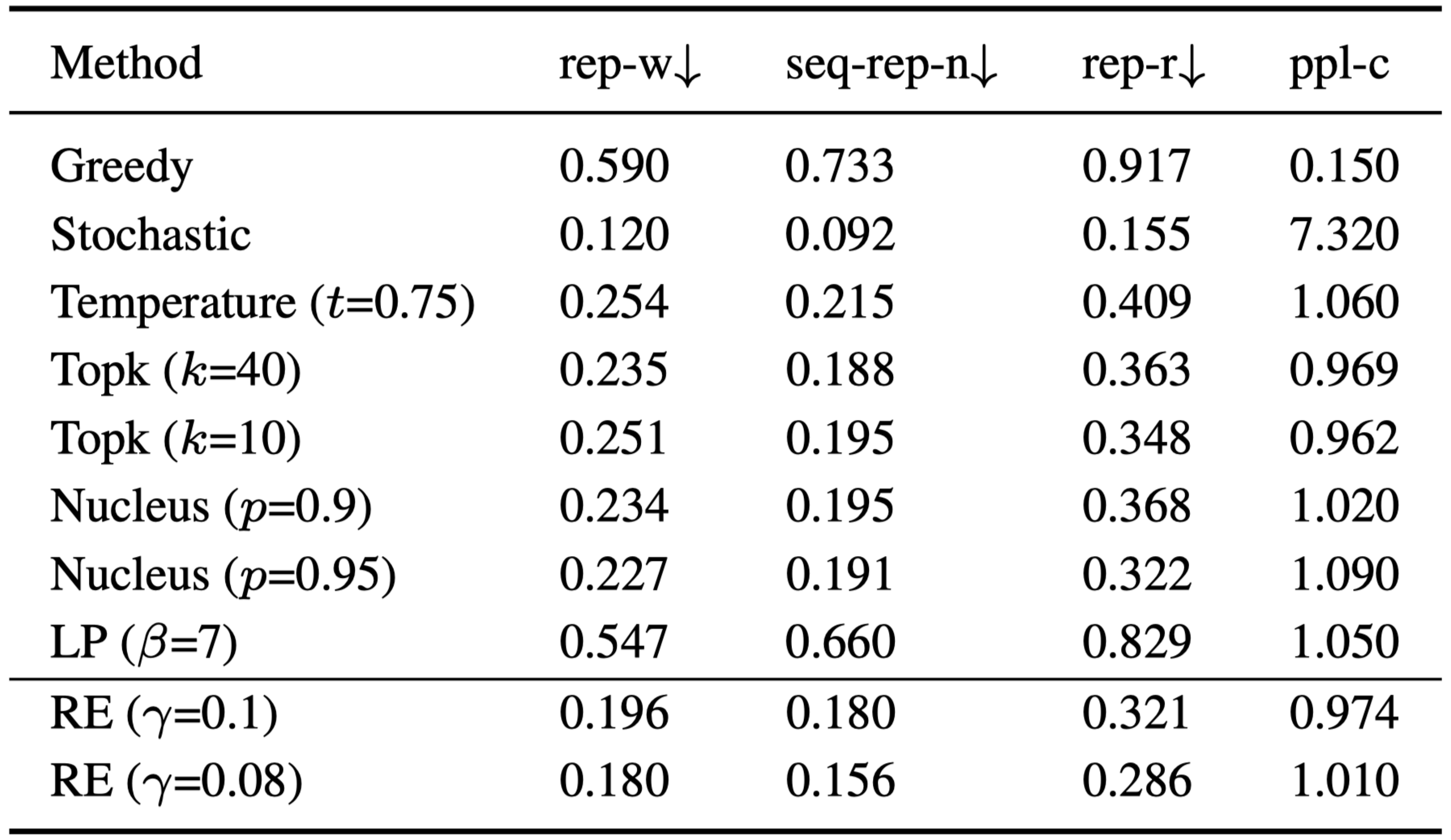
Conclusions:

- (1) The RE method alleviates the repetition problem and outperforms existing methods significantly.
- (2) The repetition problem of the Greedy method is more serious than other models while it hardly appears in the Stochastic sampling method. (Corollary 1.1)
- (3) The LP method and RE method both alleviate the repetition problem. The reason is that the RE method controls the inflow term while the LP method controls the outflow term and thus they limit the upper bound of ARP. (Corollary 1.2)



Method	rep-w↓	seq-rep-n↓	rep-r↓	BLEU↑	ROUGE$_L$ ↑
Greedy	0.0883	0.0330	0.0512	0.352	0.606
Stochastic	0.0783	0.0272	0.0337	0.222	0.472
Temperature (t=0.15)	0.0879	0.0328	0.0511	0.351	0.605
Topk (k=10)	0.0882	0.0329	0.0507	0.350	0.605
Topk (k=40)	0.0881	0.0329	0.0511	0.350	0.604
Nucleus (p=0.9)	0.0878	0.0328	0.0508	0.349	0.603
Nucleus (p=0.95)	0.0882	0.0329	0.0510	0.350	0.604
LP (β=6)	0.0863	0.0322	0.0500	0.349	0.605
RE (γ=0.15)	0.0768	0.0281	0.0419	0.350	0.608
RE (γ=0.1)	0.0743	0.0275	0.0417	0.350	0.607
RE (γ=0.05)	0.0585	0.0211	0.0296	0.340	0.603
RE (γ=0.02)	0.0434	0.0158	0.0221	0.335	0.600

Table 1: Experimental results for NMT task.[1]

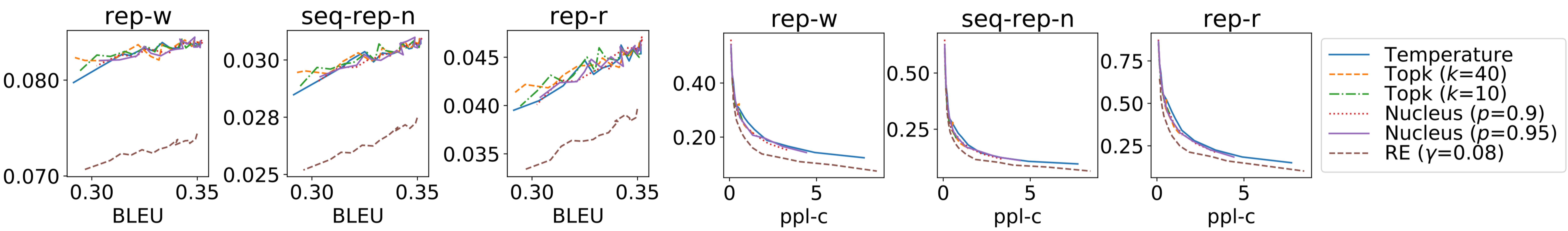


Method	rep-w↓	seq-rep-n↓	rep-r↓	ppl-c
Greedy	0.590	0.733	0.917	0.150
Stochastic	0.120	0.092	0.155	7.320
Temperature (t=0.75)	0.254	0.215	0.409	1.060
Topk (k=40)	0.235	0.188	0.363	0.969
Topk (k=10)	0.251	0.195	0.348	0.962
Nucleus (p=0.9)	0.234	0.195	0.368	1.020
Nucleus (p=0.95)	0.227	0.191	0.322	1.090
LP (β=7)	0.547	0.660	0.829	1.050
RE (γ=0.1)	0.196	0.180	0.321	0.974
RE (γ=0.08)	0.180	0.156	0.286	1.010

Table 2: Experimental results for LM task.[1]

Experiments - Repetition Performance Balancing

- Our proposed RE method achieves the best balance between the overall performance and the repetition problem.
- It can be concluded from the results that :
  - (1) The RE method achieves the lowest repetition scores at the same performance score in both NMT and LM tasks.
  - (2) The Topk sampling and Nucleus sampling alleviate the repetition problem of the Temperature sampling method because they help alleviate the long-tail effect and thus improve the overall performance. (Corollary 1.1)



(a) NMT

(b) LM

Experiments - Well-definedness of ARP

- We conduct an experiment to study the relationship between the theoretical ARP and the repetition metrics.
- To calculate the theoretical ARP, we conduct this experiment on a Markov generation model. The Markov transition matrix is calculated by counting words in Wiki-103.
- It can be concluded from the results that as the ARP grows, all repetition metrics are increasing. This positive correlation shows that ARP is well-defined.

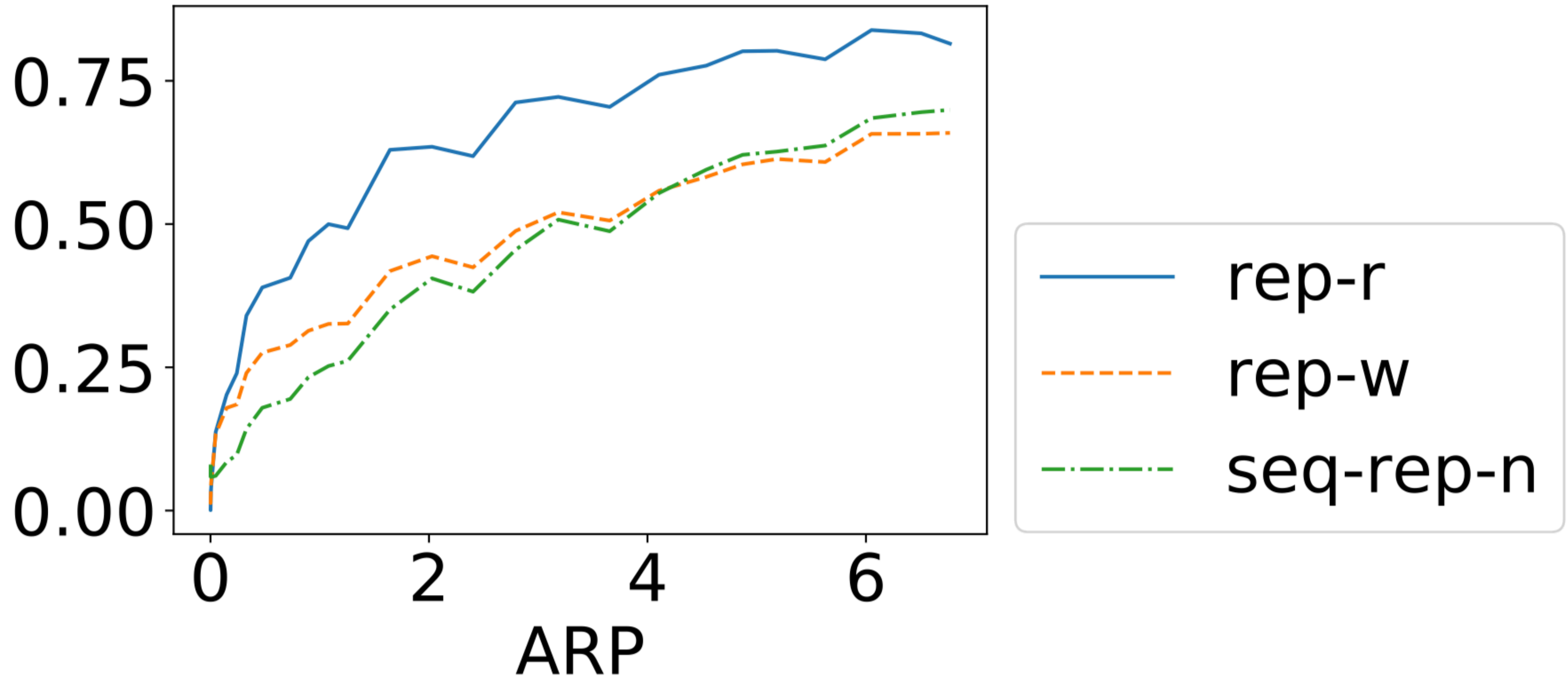


Figure 4: Correlation of ARP and repetition metrics.

Experiments - Influence of High Inflow Pairs

- To show that the high inflow pairs do cause the repetition problem, we conduct an experiment to show the relationship between the high inflow pair count and the repetition metrics.
- we calculate the high inflow pair count and the repetition scores in each sentence.
- It can be concluded from the results that if one sentence contains too many high inflow pairs, it is more likely to get a higher repetition score. (Corollary 1.2)

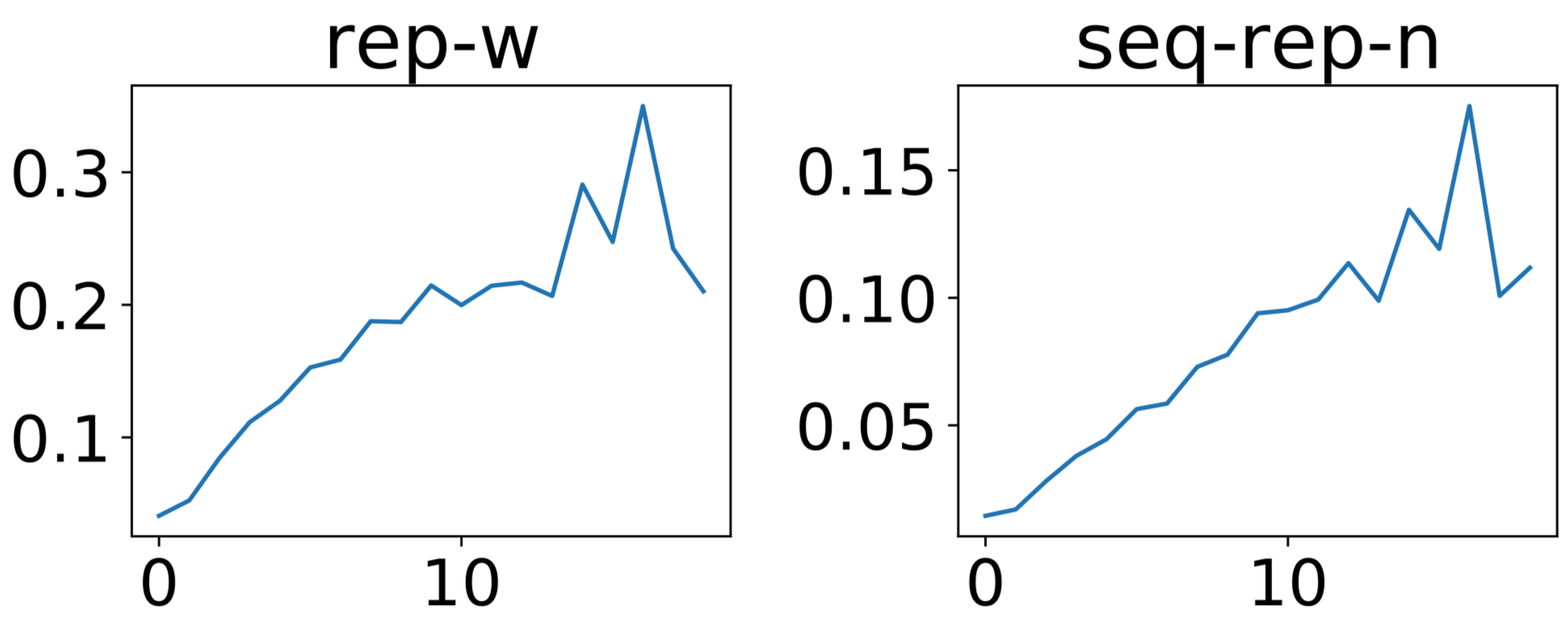
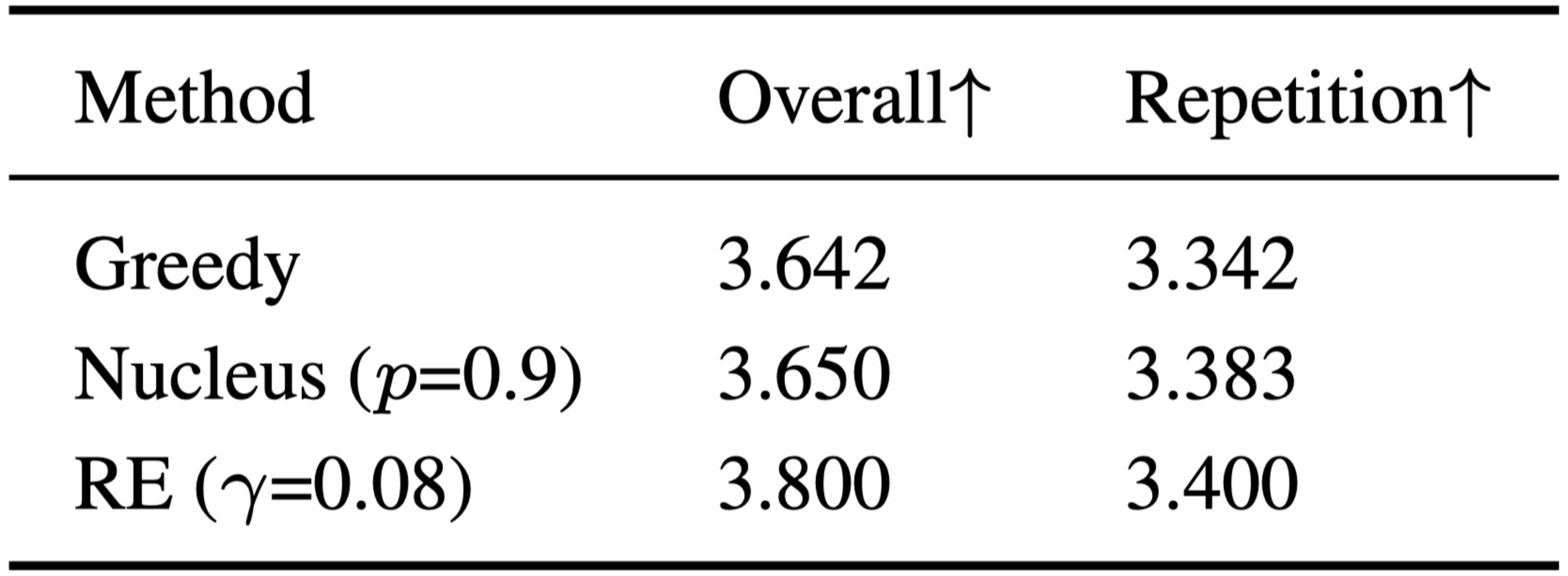


Figure 5: Influence of high inflow pair. x axis is the high inflow pair count.

Experiments - Human Evaluation

- We sample 120 generated text in the LM task for each model. The sampled text is scored by human helpers to evaluate the overall performance and repetition performance.
- Our proposed method outperforms the comparison methods in both alleviating the repetition problem and improving the overall score simultaneously.

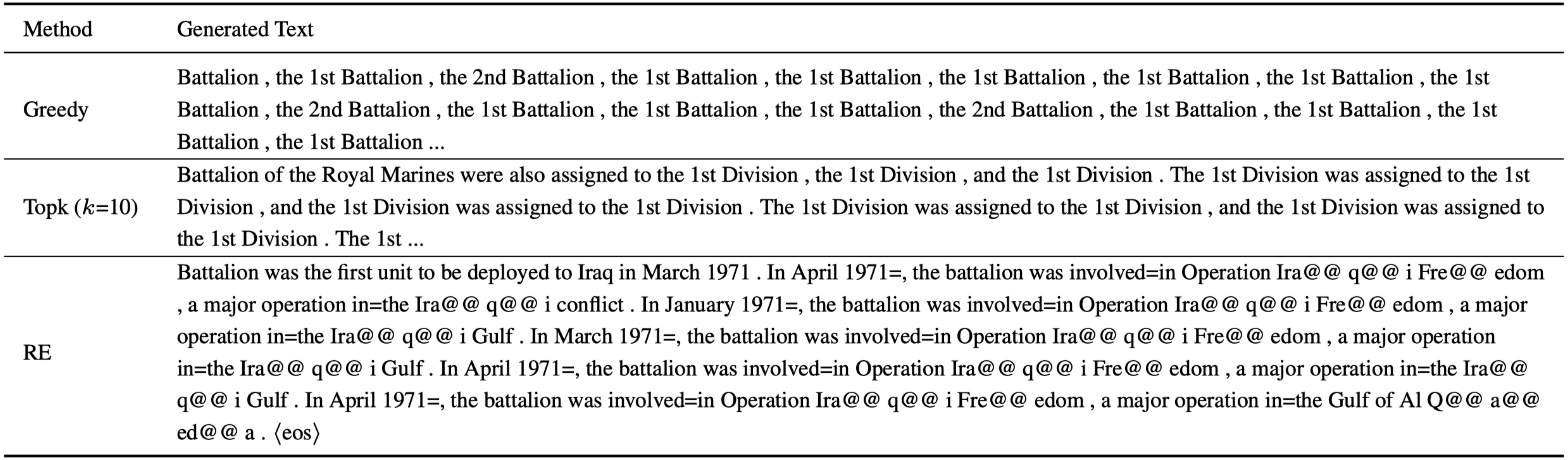


Method	Overall↑	Repetition↑
Greedy	3.642	3.342
Nucleus (p=0.9)	3.650	3.383
RE (γ=0.08)	3.800	3.400

Figure 6: Human Evaluation.[1]

Experiments - Case Study

- Greedy sampling method has a severe repetition problem.
- The Topk sampling method alleviates the repetition problem by minimizing ppl-c and can thus increase the temperature when the ppl-c level is fixed.
- Our proposed RE method gives the best trade-off between the repetition problem and the overall performance.



Method	Generated Text
Greedy	Battalion , the 1st Battalion , the 2nd Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion , the 2nd Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion , the 2nd Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion , the 1st Battalion ...
Topk (k=10)	Battalion of the Royal Marines were also assigned to the 1st Division , the 1st Division , and the 1st Division . The 1st Division was assigned to the 1st Division , and the 1st Division was assigned to the 1st Division . The 1st Division was assigned to the 1st Division , and the 1st Division was assigned to the 1st Division . The 1st ...
RE	Battalion was the first unit to be deployed to Iraq in March 1971 . In April 1971=, the battalion was involved=in Operation Ira@@ q@@ i Fre@@ edom , a major operation in=the Ira@@ q@@ i conflict . In January 1971=, the battalion was involved=in Operation Ira@@ q@@ i Fre@@ edom , a major operation in=the Ira@@ q@@ i Gulf . In March 1971=, the battalion was involved=in Operation Ira@@ q@@ i Fre@@ edom , a major operation in=the Ira@@ q@@ i Gulf . In April 1971=, the battalion was involved=in Operation Ira@@ q@@ i Fre@@ edom , a major operation in=the Ira@@ q@@ i Gulf . In April 1971=, the battalion was involved=in Operation Ira@@ q@@ i Fre@@ edom , a major operation in=the Gulf of Al Q@@ a@@ ed@@ a . ⟨eos⟩

Table 3: Experimental results for LM task.

Outline

✓Introduction

✓Theoretical Analysis

✓Rebalanced Encoding

✓Experiments

➡Conclusions

Conclusions

- **New theoretical framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More understanding of our language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified understanding of existing algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A new encoding method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.

Reference

- Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In International Conference on Learning Representations (ICLR).
- Vig, J. 2018. Deconstructing bert: Distilling 6 patterns from 100 million parameters. Medium, December .
- Choi, Y. 2018. The Missing Representation in Neural Language Models. In 3rd Workshop on Representation Learning for NLP (RepL4NLP).
- Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; and Weston, J. 2020. Neural Text Generation With Unlikelihood Training. In International Conference on Learning Representations (ICLR).

Thank You

https://github.com/fuzihaofzh/repetition-problem-nlg

1 Min Short

A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi

- Repetition Problem in text generation

Tough it is still unfinished, $\underbrace{\textbf{but I like it but I like it but I like}}_{\text{Repetition}}$ **...**

- **New Theoretical Framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More Understanding of Our Language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified Understanding of Existing Algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A New Encoding Method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.

A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi

- Repetition Problem in text generation

Tough it is still unfinished, $\underbrace{\text{but I like it but I like it but I like}}_{\text{Repetition}}$ **...**

- **New Theoretical Framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More Understanding of Our Language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified Understanding of Existing Algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A New Encoding Method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.

A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi

- Repetition Problem in text generation

Tough it is still unfinished, but I like it but I like it but I like ...

Repetition

- **New Theoretical Framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More Understanding of Our Language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified Understanding of Existing Algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A New Encoding Method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.

A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi

- Repetition Problem in text generation

Tough it is still unfinished, $\underbrace{\textbf{but I like it but I like it but I like}}_{\text{Repetition}}$ ...

- **New Theoretical Framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More Understanding of Our Language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified Understanding of Existing Algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A New Encoding Method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.

A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi

- Repetition Problem in text generation

Tough it is still unfinished, $\underbrace{\text{but I like it but I like it but I like}}_{\text{Repetition}}$ **...**

- **New Theoretical Framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More Understanding of Our Language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified Understanding of Existing Algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A New Encoding Method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.

A Theoretical Analysis of the Repetition Problem in Text Generation

Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi

- Repetition Problem in text generation

Tough it is still unfinished, $\underbrace{\textbf{but I like it but I like it but I like}}_{\text{Repetition}}$ **...**

- **New Theoretical Framework**
  - We propose a novel theoretical framework to analysis the repetition problem.
- **More Understanding of Our Language**
  - Our theory show that the repetition problem is, unfortunately, caused by our language itself. There exist too many words predicting the same word as the subsequent word with high probability. Consequently, it is easy to go back to that word and form repetitions and we dub it as the high inflow problem.
- **Unified Understanding of Existing Algorithms**
  - Our theory provide a unified understanding of existing methods alleviating the repetition problem, including stochastic sampling, topk sampling, nucleus sampling, temperature sampling, length penalty, and etc.
- **A New Encoding Method**
  - Based on the theory, we propose a novel rebalanced encoding approach to solve the high inflow problem and thus alleviate the repetition problem.