Diverse Keyphrase Generation with Neural Unlikelihood Training

Problem Formulation

- Supervised Setting
- **Input**: a source document (e.g. a news article)
- **Output**: a set of keyphrases that describe the main ideas presented in the source document

Cathay Pacific Airlines Fined Over Data Breach

The U.K. Information Commissioner's Office has fined Cathay Pacific Airways £500,000 ($646,000) over a data breach that exposed the personal information of 9.4 million customers, including 111,000 British citizens, during a four-year period.

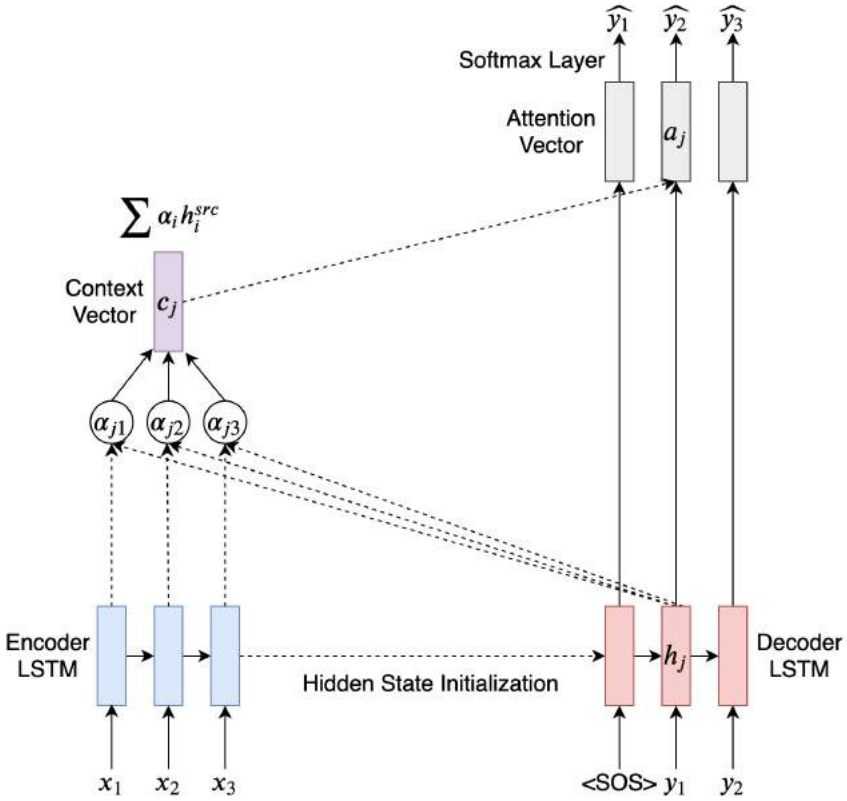
See Also: Maintaining Continuous Compliance with GDPR

The fine is the largest the U.K. privacy watchdog could impose under the country's older data protection laws since the breach, which started in 2014 and was discovered and fixed in 2018. That happened before the EU's General Data Protection Regulation went into effect in May 2018, according to the report.

During its investigation, the ICO found that the Hong Kong-based airline lacked appropriate security controls to ensure passenger data was secured within its internal IT systems, according to the report. The result is that millions of records, including names, passport and

cathay pacific airline | data breach | personal information leaked | fines | privacy protection

Sequence-to-Sequence Attention Models



Incorporating Copy Mechanism

- At each decoding time-step, predict [p_{target} , p_{copy}]
  - p_{target} * [target_vocabulary_prob_scores]
  - p_{copy} * [source_token_prob_scores]
  - Re-normalize probabilities and predict the most probable word

Maximum Likelihood Training

- Decoder generates output token by token conditioned on the input and the previous words in the context

$$\mathcal{L}_{\text{MLE}} = -\sum_{t=1}^{L} \log P(y_t | \mathbf{y}_{1:t-1}, \mathbf{x}, \boldsymbol{\theta})$$

MLE Training Issues

- Generation of redundant keyphrases
- Qualitative Example
- ***Quantitative*** ways to measure diversity?

Title	semi automated schema integration with sasmint
Abstract	the emergence of increasing number of collaborating organizations has made clear the need for supporting interoperability infrastructures , enabling sharing and exchange of data among organizations . schema matching and schema integration are the crucial components of the interoperability infrastructures , and their semi automation to interrelate or integrate heterogeneous and autonomous databases in collaborative networks is desired . the semi automatic schema matching and integration sasmint system introduced in this paper identifies and resolves (...)
Ground Truth	*schema integration ; collaboration ; schema matching ; heterogeneity ; data sharing*
MLE Baseline	*schema integration ; sasmint ; schema matching ; schema integration ; schema matching ; sasmint derivation markup language*

MLE Training Issues

- Generation of redundant keyphrases
- # of keyphrases in generated is different from ground truth
- % of repeated tokens and keyphrases (ground truth vs. model generated) - after stemming

	#Keyphrases	% duplicate keyphrases	% duplicate tokens
Ground Truth	5.3	0.1	7.3
MLE Baseline	7.3	26.6	36.0

Adhoc Post-processing Solutions

- Exhaustive beam search decoding
- Generate large number of keyphrases and then prune/de-duplicate
  - computationally expensive
  - wasteful because only < 5% of such KPs are unique
- More of a hacky last-minute solution. Not a principled approach
- We would like to ideally address this redundancy issue directly at training time

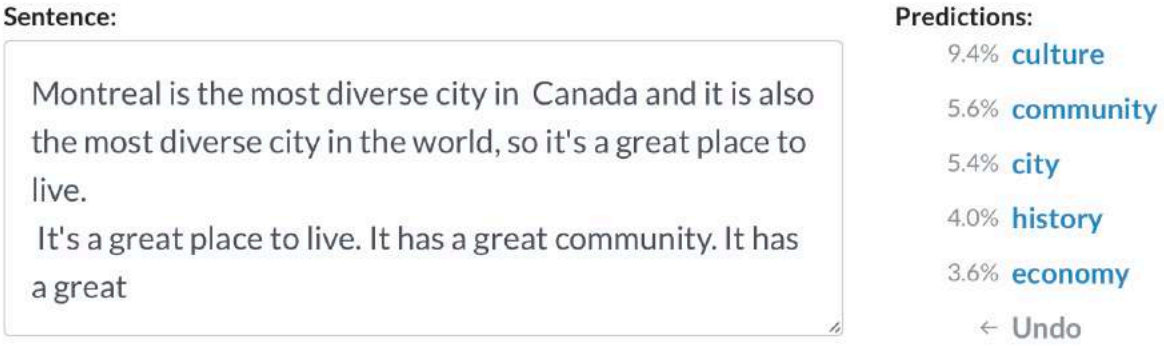
Proposed Solution 1
Unlikelihood Training

Motivation

- MLE optimizes the likelihood of the entire data distribution:
    - not focussed on optimizing the current sequence being generated
- Likelihood training assigns too much probability to sequences containing repeats and frequent words
- **Unlikelihood training:**
    - Do regular likelihood update on true target tokens
    - Unlikelihood update (penalizing) on tokens that are *"unnecessarily"* assigned high probability

Neural Text Generation with Unlikelihood Training (Welleck et. al, ICLR 2020).

Example

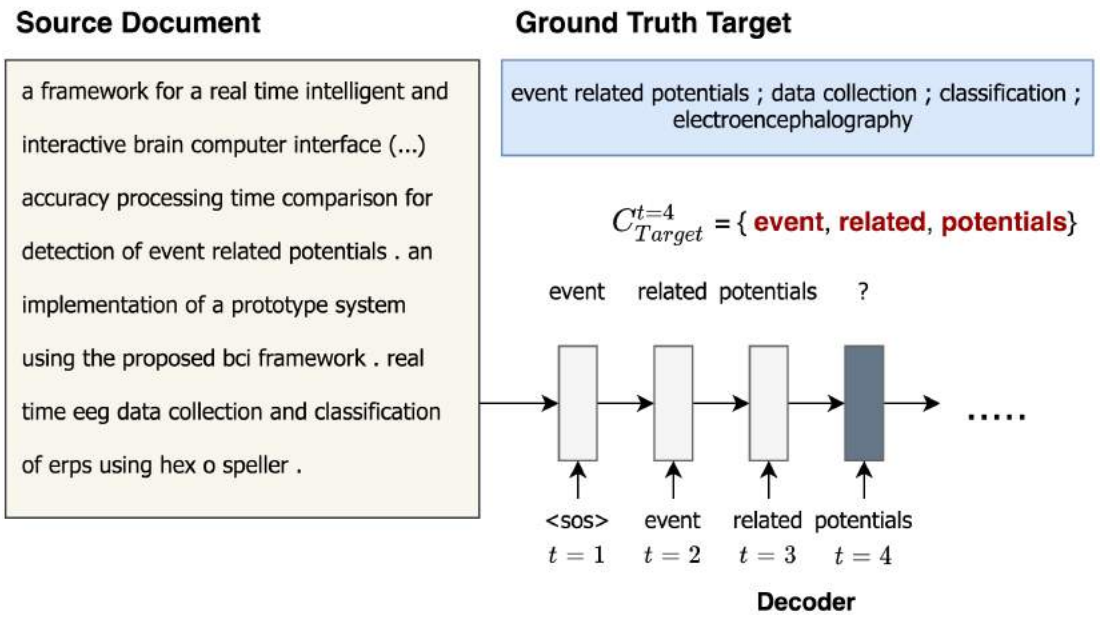


- With MLE training, generations can become quite mundane and un-interesting!
- Unlikelihood paper —> demonstrate method for language modelling

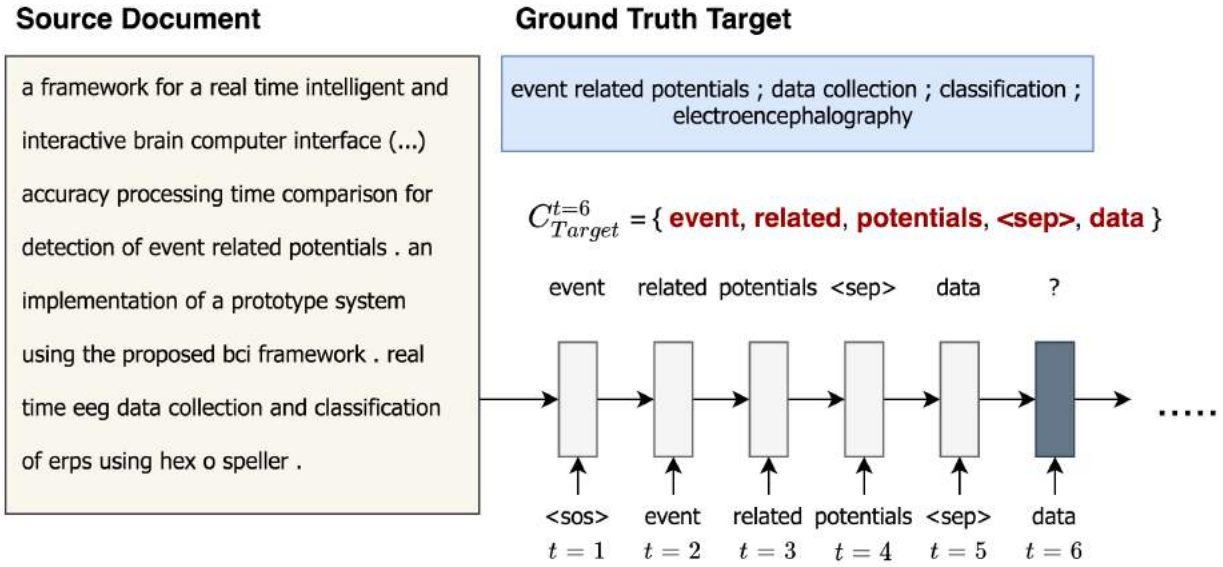
Source: https://demo.allennlp.org/next-token-lm

Token Level Unlikelihood Training

- During generation process, keep a **negative candidate list**
- Penalize when high probability is assigned to words in **negative candidate list**



Target Token Level Unlikelihood Training

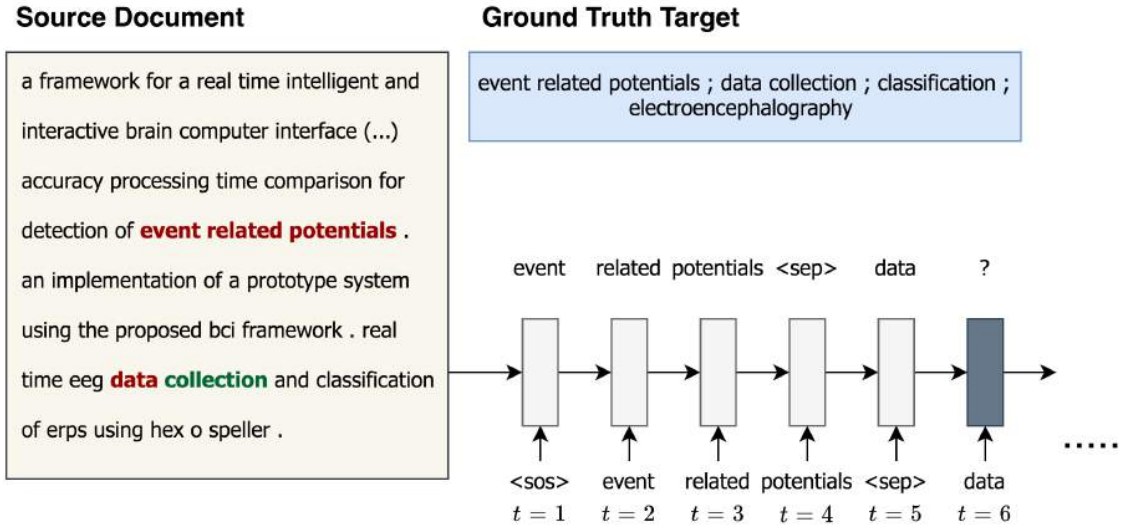


$$\mathcal{L}_{\text{TargetUL}} = -\sum_{t=1}^{L} \sum_{c \in \mathcal{C}_{\text{Target}}^{t}} \log\left(1 - P_{target}(c|\mathbf{y}_{1:t-1}, \mathbf{x}, \boldsymbol{\theta})\right)$$

- Our negative candidate list consists of the ground truth tokens from the previous time steps
- Target probabilities used to compute loss
- This way the model is penalized for repeatedly generating the same KP over and over again

Copy Token Level Unlikelihood Training

- Remember that the model can also copy from the source text during decoding
- Copy UL loss discourages the behaviour of repeatedly copying the same set of words from the source



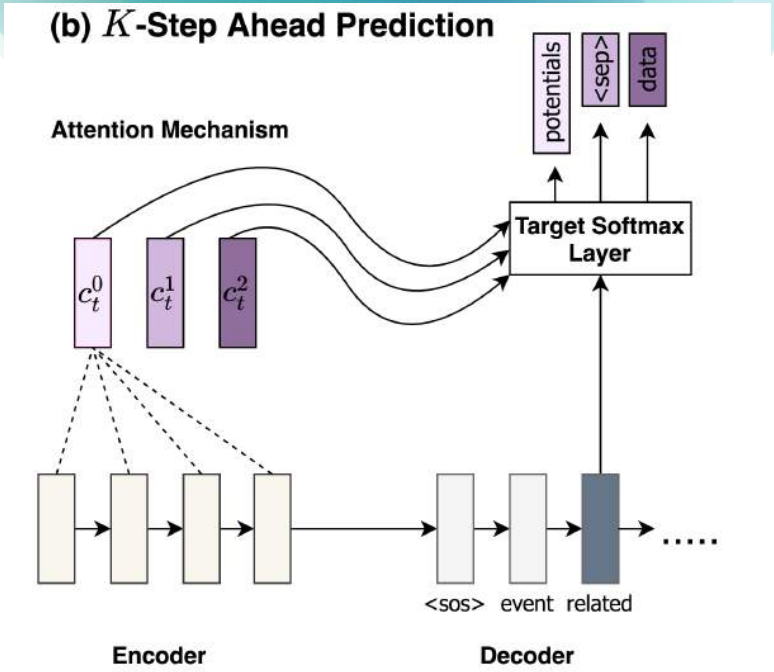
- **Negative candidate list :** ground truth context tokens from previous time steps that also appear in the source text (and thus can be copied)
- Loss is computed based on copy probabilities

$$\mathcal{L}_{\text{CopyUL}} = -\sum_{t=1}^{L} \sum_{c \in \mathcal{C}^t_{\text{Copy}}} \log\left(1 - P_{copy}(c|\mathbf{y}_{1:t-1}, \mathbf{x}, \boldsymbol{\theta})\right)$$

Proposed Solution 2
K-Step Ahead Token Prediction

K-Step Ahead MLE

- Regular S2S objective is based on the next token prediction
- Greedy approach - does not incentivize the model to plan for the upcoming future tokens ahead of time
- Ask the model to simultaneously predict future tokens until K-steps ahead
- Use the same decoder hidden state, but learn **different attention mechanisms** over the source tokens for k=0 to K



$$\mathcal{L}_{K-\text{StepMLE}} = -\sum_{t=1}^{L}\sum_{k=0}^{K} \gamma_k \log P(y_{t+k}|\mathbf{y}_{1:t-1}, \mathbf{x}, \boldsymbol{\theta})$$

K-Step Ahead UL Losses

- We saw MLE-based loss for the task of K-step ahead token prediction.
- Can be naturally extended to the unlikelihood setting
- Impose the target and copy unlikelihood losses on the K-step ahead token prediction task
- Should further improve diversity

$$\mathcal{L}_{K-\text{StepTargetUL}} = -\sum_{t=1}^{L}\sum_{k=0}^{K}\gamma_k \sum_{c \in \mathcal{C}_{\text{Target}}^{t+k}} \log\left(1 - P_{target}(c|\mathbf{y}_{1:t-1}, \mathbf{x}, \boldsymbol{\theta})\right)$$

$$\mathcal{L}_{K-\text{StepCopyUL}} = -\sum_{t=1}^{L}\sum_{k=0}^{K}\gamma_k \sum_{c \in \mathcal{C}_{\text{Copy}}^{t+k}} \log\left(1 - P_{copy}(c|\mathbf{y}_{1:t-1}, \mathbf{x}, \boldsymbol{\theta})\right)$$

Overall Training Objective

- We combine the losses:

$$\mathcal{L} = \mathcal{L}_{K-\text{StepMLE}} + \lambda_T \mathcal{L}_{K-\text{StepTargetUL}} + \lambda_C \mathcal{L}_{K-\text{StepCopyUL}}$$

Datasets , Baselines and Evaluation Metrics

Datasets

- **KP20K**: Scientific article abstracts and associated keyphrases
- **KPTimes**: News articles and editor-assigned keyphrases
- **StackExchange**: Community QA forum with question description and user assigned tags

Baselines

- **catSeq**: A S2S model trained solely using the MLE objective
- **catSeqD** (Yuan et. al, 2020): auxiliary semantic coverage and orthogonality losses to enhance generation diversity
- **catSeqCorr** (Chen et al. (2018): catSeq model with a coverage module and review mechanism
- **catSeqTG** (Chen et al. (2019): separately encodes the title information using an attention-guided matching layer
- **catSeqTG-2RF1** (Chan et al. (2019): catSeqTG model + reinforcement learning objective where F_1-score is the training reward

Quality/Relevance Metrics

- Obtained by comparing with ground truth KPs
- Precision - how many correct KPs among all the model generated KPs ?
- Recall - how many of the ground truth KPs were generated by the model ?
- F1-score

Diversity Metrics

- % Duplicate KPs
- % Duplicate Tokens
- # KPs
- Inter-keyphrase similarity among the generated set of keyphrases - a lower value indicates fewer repetitions and thus more diversity in the output.
  - **Self-BLEU**: Compute pairwise BLEU score between generated KPs; captures word level surface overlap.
  - **EditDist**: Character level string matching; pairwise Levenshtein Distance between generated KPs
  - **EmbSim**:
    - With Self-BLEU and EditDist, we can only capture surface level repetitions between KPs
    - Use pre-trained phrase-level embeddings that measures inter-keyphrase similarity at a semantic level
    - compute pairwise cosine similarities between Sent2Vec embedding representations of keyphrases

Results

Quantitative Results

		Quality Evaluation			Diversity Evaluation					
		$P@M$	$R@M$	$F_1@M$	#KPs	%Duplicate KPs ↓	%Duplicate Tokens ↓	Self-BLEU ↓	Edit-Dist ↓	Emb-Sim ↓
Scientific Articles - KP20K	Ground Truth	-	-	-	→5.3	0.1	7.3	3.8	32.7	0.159
	catSeq	0.291	0.260	0.274	7.3	26.6	36.0	26.6	45.6	0.328
	catSeqD	0.294	0.257	0.274	6.7	25.7	35.3	27.0	45.3	0.325
	catSeqCorr	0.283	0.264	0.273	7.0	23.2	33.5	24.5	44.0	0.309
	catSeqTG	**0.295**	0.262	0.278	6.8	24.7	34.3	26.2	45.2	0.323
	catSeqTG-2RF1	0.274	**0.286**	**0.280**	7.5	30.9	41.7	30.7	46.7	0.341
	DivKGen (UL)	0.277	0.261	0.269	**5.0**	5.3	12.6	9.7	**34.4**	**0.181**
	+K-StepMLE	0.274	0.239	0.255	4.6	6.1	13.9	11.5	36.2	0.197
	+K-StepUL	0.273	0.240	0.256	4.6	**4.9**	**11.7**	**8.8**	35.2	0.185

- All MLE models have high percentage of repetitions
- RL based model achieves best F1, but worst in terms of diversity

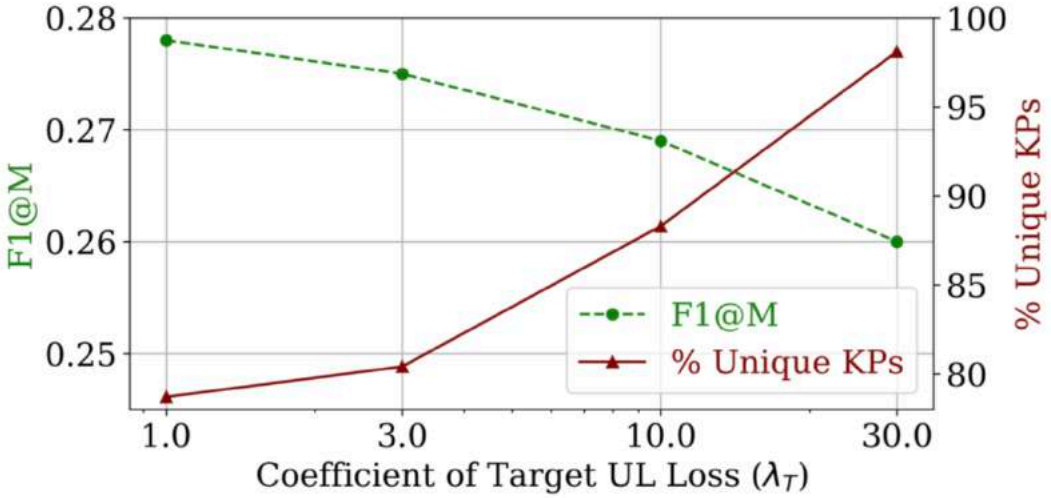
Ablation Study

DivKGen Variants	Overall $F_1@M \uparrow$	%Duplicate KPs ↓	%Duplicate Tokens ↓	Self-BLEU↓
w/ TargetUL	**0.277**	12.0	19.8	16.7
w/ CopyUL	0.263	14.1	22.7	19.9
w/ K-StepMLE	0.265	12.6	18.9	16.3
w/ TargetUL + CopyUL	0.269	5.3	12.6	9.7
+ K-StepMLE	0.255	6.1	13.9	11.5
+ K-StepUL	0.256	**4.9**	**11.7**	**8.8**

- Each individual loss component is not as effective as their combination
- Losses contribute in a synergetic manner to maximize diversity gains

Diversity/Quality Trade-off

- Clear quality-diversity trade-off
- Tune diversity hyperparameter as desired



Qualitative Results

Dataset : KP20K	
Title	automatic image segmentation by dynamic region merging .
Abstract	this paper addresses the automatic image segmentation problem in a region merging style . with an initially oversegmented image , in which many regions or superpixels with homogeneous color are detected , an image segmentation is performed by iteratively merging the regions according to a statistical test . there are two essential issues in a region merging algorithm order of merging and the stopping criterion . in the proposed algorithm , these two issues are solved by a novel predicate , which is defined by the sequential probability ratio test and the minimal cost criterion . starting from an oversegmented image , neighboring regions are progressively merged if there is an evidence for merging according to this predicate . we show that the merging order follows the principle of dynamic programming . this formulates the image segmentation as an inference problem , where the final segmentation is established based on the observed image . we also prove that the produced segmentation satisfies certain global properties . in addition , a faster algorithm is developed to accelerate the region merging process , which maintains a nearest neighbor graph in each iteration . experiments on real natural images are conducted to demonstrate the performance of the proposed dynamic region merging algorithm .
Ground Truth	*image segmentation ; region merging ; dynamic programming ; wald sequential probability ratio test*
catSeq MLE Baseline	*image segmentation ; region merging ; region merging ; dynamic programming ; image segmentation*
catSeqTG-2RF1 (RL)	*image segmentation ; region merging ; dynamic programming ; image segmentation ; dynamic programming*
DivKGen (UL)	*image segmentation ; region merging ; region merging ; dynamic programming ; nearest neighbor graph*
DivKGen (Full)	*image segmentation ; dynamic programming ; region merging ; stopping criterion*

Summary

- Shortcoming of MLE based training for KP Generation
- Unlikelihood training
    - —> more diversity ; much lower repetitions
    - —> closer to actual data distributions
- K-Step ahead losses to improve model planning capability
- Diversity / Quality trade-off

Thank You

Questions?