A Comparative Study of Training Objectives for Clarification Facet Generation

Shiyu Ni, Keping Bi, Jiafeng Guo, Xueqi Cheng

CAS Key Lab of Network Data Science and Technology, ICT, CAS, Beijing, China

University of Chinese Academy of Sciences, Beijing, China

MOTIVATION

- Since user queries can be ambiguous or vague, query intent clarification is beneficial to enhance user experience and retrieval effectiveness
- Current work on query facet generation only considers whether the training objective is permutation-invariant and the evaluation is inadequate
- It is necessary to conduct a systematic comparative study of various types of training objectives with different properties

VARIOUS TRAINING OBJECTIVES

Model	Sequential-Prediction	Permutation-Invariant	Facet-Count-Controllable	Complexity-Per-Training-Epoch
Seq-Default	✓	×	×	$O(n*((\lvert q\rvert+\lvert d\rvert)^2+(m*\lvert f\rvert)^2+(\lvert q\rvert+\lvert d\rvert)*m*\lvert f\rvert))$
Seq-Min-Perm	✓	✓	×	$O(n*m!*((\lvert q\rvert+\lvert d\rvert)^2+(m*\lvert f\rvert)^2+(\lvert q\rvert+\lvert d\rvert)*m*\lvert f\rvert))$
Seq-Avg-Perm	✓	✓	×	$O(n*m!*((\lvert q\rvert+\lvert d\rvert)^2+(m*\lvert f\rvert)^2+(\lvert q\rvert+\lvert d\rvert)*m*\lvert f\rvert))$
Set-Pred	×	✓	✓	$O(n*m*((\lvert q\rvert+\lvert d\rvert)^2+\lvert f\rvert^2+(\lvert q\rvert+\lvert d\rvert)*\lvert f\rvert))$
Seq-Set-Pred	✓	✓	✓	$O(n*(\sum_{i=0}^{m-1}\mathcal{A}_m^i*\mathcal{A}_{m-i}^1)*((\lvert q\rvert+\lvert d\rvert+i*\lvert f\rvert)^2+\lvert f\rvert^2+(\lvert q\rvert+\lvert d\rvert+i*\lvert f\rvert)*\lvert f\rvert))$

Setting of Query Facet Generation

Let $D=\{(q_1,D_1,F_1),(q_2,D_2,F_2),\cdots,(q_n,D_n,F_n)\}$ denote the training data , where q_i is the i - th open-domain query, $D_i=\{d_{i1},d_{i2},\cdots,d_{ik}\}$ is the top k_i retrieved documents for the given query. $F_i=\{f_{i1},f_{i2},\cdots,f_{im}\}$ represents m_i ground truth facets related to q_i. The task is to generate a set of related facets F for any given query q with its associated documents D.

- **Seq-Avg-Perm**
  - Sequentially generates facets and is trained with the average loss of permutations of facets

$$\theta=\arg\min_{\theta^*}\sum_{i=1}^{n}\frac{1}{\lvert\pi(y_i)\rvert}\sum_{y_i^*\in\pi(y_i)}P(v,y_i^*)$$

  Where $P(v,y_i)=\frac{1}{\lvert y_i\rvert}\sum_{x=1}^{\lvert y_i\rvert}-\log p(y_{ix}\mid v,y_{i1},\cdots,y_{ix-1})$, v is the output of the encoder, and $\pi(y_i)$ means all the permutations

  - We permute the concatenation of facets in order to let the model learn towards all the possible permutations
- **Set-pred**
  - Treats each facet as an individual target and conducts parallel predictions

$$\theta=\arg\min_{\theta^*}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{f_{ij}\in F_i}P(v,f_{ij})$$

  where P and v are the same as seq-avg-perm.

  - Do not use the so-far predicted facets as contexts for the current facet generation and the generated facets may be synonyms
- **Seq-Set-Pred**
  - Predicts each of the remaining facet sets in parallel based on arbitrarily generated facets as context

$$\theta=\arg\min_{\theta^*}\sum_{i=1}^{n}\frac{1}{V_i}\sum_{j=0}^{\lvert F_i\rvert-1}\sum_{k=1}^{\lvert\mathcal{A}_j(F_i)\rvert}\sum_{h=1}^{\lvert\mathcal{A}_1(R_{ijk})\rvert}P(v_{ijk},y_{ijkh})$$

  where $\mathcal{A}_j(X)$ means all the possible orderings of j selected non-repetitive elements from X

  - It is a combination of seq-avg-perm and set-pred

EXPERIMENTS

- **Model and Dataset**

Out dataset is MIMICS and the model is BART-base

  - Dataset: MIMICS is a collection of search clarification datasets for real search queries sampled from the Bing query logs
  - Model: BART-base is a Seq2Seq model for sequence generation tasks

- **Newly Proposed Metrics**

We introduce two metrics to measure the diversity of generated facets

  - Term diversity: calculate the average of one minus the overlap ratio between each pair of facets
  - BERT-Score diversity: calculate the average BERTScore between each pair of facets

- **Evaluation Against Ground Truth**
  - Most permutation-invariant methods except seq-min-perm perform better than the order-sensitive methods
  - The method that generates facets without depending on the previously generated facets (i.e., set-pred) has compelling performance in terms of both term-based and semantic matching metric
  - Methods that only learn facet prediction given context(e.g., seq-set-pred) have better semantic matching performance with ground truth compared to those that also learn when to stop generating facets

Model	Term Overlap			Exact Match			Set BLEU Score				Set BERT-Score		
	P	R	F1	P	R	F1	1-gram	2-gram	3-gram	4-gram	P	R	F1
Seq-Default	0.2976^{+}	0.2769^{+-}	0.2752^{+-}	0.0718^{-}	0.0561^{-}	0.0611^{-}	0.2335^{+-}	0.1040^{+-}	0.0444^{+-}	0.0175^{+-}	0.6391^{-}	0.6455^{-}	0.6419^{-}
Seq-Min-Perm	0.2761^{-}	0.2536^{-}	0.2537^{-}	0.0620^{-}	0.0470^{-}	0.0519^{-}	0.2102^{-}	0.0850^{-}	0.0346^{-}	0.0125^{-}	0.6442^{+-}	0.6482^{-}	0.6457^{+-}
Seq-Avg-Perm	0.2977^{+}	$\mathbf{0.3263}^{+}$	$\mathbf{0.3005}^{+}$	$\mathbf{0.1040}^{+}$	0.0960^{+}	$\mathbf{0.0977}^{+}$	0.2422^{+-}	0.1081^{+-}	0.0568^{+-}	0.0288^{+}	0.6665^{+-}	0.6697^{+-}	0.6676^{+-}
Set-Pred	$\mathbf{0.3029}^{+}$	0.2978^{+-}	0.2897^{+}	0.0988^{+}	0.0973^{+}	0.0953^{+}	0.2567^{+}	0.1198^{+}	0.0606^{+-}	0.0260^{+-}	$\mathbf{0.6873}^{+}$	$\mathbf{0.6897}^{+}$	$\mathbf{0.6880}^{+}$
Seq-Set-Pred	0.2930^{+}	0.2989^{+-}	0.2863^{+-}	0.0993^{+}	$\mathbf{0.1009}^{+}$	0.0973^{+}	$\mathbf{0.2577}^{+}$	$\mathbf{0.1228}^{+}$	$\mathbf{0.0676}^{+}$	$\mathbf{0.0308}^{+}$	0.6849^{+}	0.6887^{+}	0.6863^{+}

- **Evaluation on Diversity**
  - Seq-avg-perm performs the best on both metrics
  - All the sequential prediction methods, except for seq-default, exhibit higher diversity than set-pred
  - We observe good term diversity but worse semantic diversity in seq-set-pred

Model	Term Diversity Ratio	BERT-Score Diversity P	R	F1
Ground Truth	0.9284	0.0829	0.0829	0.0836
Seq-Default	0.8783	0.0743	0.0743	0.0749
Seq-Min-Perm	0.8922	0.0883	0.0889	0.0893
Seq-Avg-Perm	**0.9218**	**0.0993**	**0.0989**	**0.1000**
Set-Pred	0.8883	0.0630	0.0630	0.0635
Seq-Set-Pred	0.9117	0.0657	0.0666	0.0667

Model	Num	Term Overlap F1	Set BERT-Score F1	Term Div Ratio	BERT-Score Div F1
Set-Pred	1	0.2696	0.3254	-	-
	2	0.2888	0.6477	0.8812	0.0613
	3	**0.2897**	**0.6880**	0.8883	0.0635
	4	0.2867	0.6231	**0.8903**	0.0658
	5	0.2814	0.5450	0.8869	**0.0683**
Seq-Set-Pred	1	0.2709	0.3253	-	-
	2	**0.2886**	0.6474	**0.9133**	0.0649
	3	0.2863	**0.6863**	0.9117	0.0667
	4	0.2759	0.6259	0.9095	0.0687
	5	0.2677	0.5660	0.9083	**0.0705**

- **Performance w.r.t. Facet Counts**
  - The best matching scores are mainly achieved when generating 2 or 3 facets
  - Compared to seq-avg-perm, facet-count-controllable methods perform better on set BLEU score and set BERT-Score
  - Seq-set-pred demonstrates better diversity than set-pred across all the numbers of generated facets.

	Seq-default	Seq-Min-Perm	Seq-Avg-Perm	Three Facets	Two Facets
Ratio	0.7038	0.7072	0.6678	0.7039	**0.7431**

- **Impact of Training Data Amount**
  - The results have different extents of regressions
  - Both seq-avg-perm and seq-set-pred still outperform seq-default and seq-min-perm in terms of all the metrics

Model	Term Overlap F1	Exact Match F1	Set BERT-Score F1
Seq-Default	0.2752^{+-}	0.0611^{+-}	0.6419^{-}
Seq-Min-Perm	0.2537^{-}	0.0519^{-}	0.6457^{+-}
Seq-Avg-Perm	$\mathbf{0.2898}^{+}$	0.0737^{+-}	0.6562^{+-}
Set-Pred	0.2897^{+}	$\mathbf{0.0953}^{+}$	$\mathbf{0.6880}^{+}$
Seq-Set-Pred	0.2777^{+-}	0.0816^{+-}	0.6791^{+-}

- **Facet Generation with ChatGPT**
  - ChatGPT has a large performance gap compared to most of the methods presented in our paper across all metrics
  - The facets generated by ChatGPT are general concepts

Model	Term Overlap F1	Exact Match F1	Set BERT-Score F1
Seq-Default	0.2566	0.0374	0.6070
Seq-Min-Perm	0.2627	0.0044	0.6293
Seq-Avg-Perm	**0.3276**	0.0760	0.6654
Set-Pred	0.3187	**0.1225**	**0.7039**
Seq-Set-Pred	0.2998	0.1003	0.5811
ChatGPT	0.1598	0.0315	0.5711

CONCLUSION

Sequential-Prediction	Permutation-Invariant	Facet-Count-Controllable
The method does not sequentially predict facets has compelling matching scores but the worst diversity	Appropriate permutation-invariant objectives can help generate better facets	Methods that only learn facet prediction have better semantic matching metrics but worse diversity