

PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji and Xueqi Cheng



1. CAS Key Lab of Network Data Science and Technology, Institute of **Computing Technology, Chinese Academy of Sciences**

2. University of Chinese Academy of Sciences



New Paradigm of NLP

- Pre-training and then fine-tuning paradigm
- Significant benefit for tasks with limited training data







Models for downstream tasks

NLP Tasks

Machine Translation

Sentiment Analysis

Question Answering

Dialogue & Chatbot

Textual Entailment

Paraphrasing

Semantic Parsing

...



• BERT: Bidirectional Encoder Representations from Transformers



- Pre-trained with mask language model and next sentence A comparison of BERT with previous SOTA on GLUE, SQUAD • \bullet prediction on Wikipedia and BookCorpus. 1.1, SQUAD 2.0, from Devlin et.al.



BERT outperform previous SOTA on many natural language understanding tasks.

BERT for Information Retrieval

• Directly applying BERT to IR



- Usage of BERT for IR. Concatenate query and document, A comparison of BERT with BM25 and previous SOTA on \bullet \bullet take [CLS] for relevance computation downstream IR tasks.



Pre-trained models also benefit the search tasks, but not very significant

Observation

•

- Pegasus for Abstractive Summarization
- SSPT for Question Answering
- SentiLARE for sentiment analysis
- ERNIE (THU) for entity-related tasks

The pre-training objective that more closely resembles the

downstream tasks leads to better and faster fine-tuning performance.

Pegasus for Abstractive Summarization

• Gap sentence generation (GSG): selected by ROUGE scores



- One sentence is masked with [MASK1] and used as target A comparison of PEGASUS with other pretrained models on ${ \bullet }$ XSum, CNN/DailyMail and Gigaword. generation text (GSG).
- Pegasus is significantly better than other pre-trained models

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, ICML, 2020

SSPT for Question Answering

entity, pseudo answer), jointly pre-training with MLM



- A comparison of BERT+SSPT with BERT on SQUAD 1.1, SQUAD Span selection training instance generation. Masked span \bullet will be predicted by the passage containing it. 2.0, HotpotQA and Natural Questions.
- BERT+SSPT is significantly better than original BERT

Span Selection Pre-training for Question Answering, ACL, 2020

• Span Selection PreTraining (SSPT): predict masked span(noun phrase or

However, pre-training objectives tailored for ad-hoc retrieval have not been well explored.

Revisit the Pre-training Objectives

Sequence-based tasks:

- Masked Language Modeling
- Permuted Language Modeling

Learn contextual representations

Sequence pair-based tasks:

- Next Sentence Prediction
- Sentence Order Prediction

Learn inter-sequence coherence

IR requirements

Good representations for the query and the document

Relevance matching between short queries and long documents

- x sentence-pair vs. query-document
- x coherence vs. relevance

Pre-training for Passage Retrieval in openQA

• Design three pre-training tasks that resemble the relevance relationship between natural language questions and answer passages

Body First Selection(BFS)

offrey Everest Hinton CCERSERSC[11] (born 6 December 1947) is an English Canadian cognitive psychologis he divides his time working for Google

the backpropagation algorithm for training multi-layer neural networks, [14] although they were not the first to propo approach.^[15] Hinton is viewed by some as a leading figure in the deep learning community and is referred to by s "Godfather of Deep Learning". [16][17][18][19][20] The dramatic image-recognition milestone of the AlexNet designed Alex Krizhevsky^[21] for the ImageNet challenge 2012^[22] helped to revolutionize the field of computer vision.^[23] Hil awarded the 2018 Turing Prize alongside Yoshua Bengio and Yann LeCun for their work on deep learning.^[24]

Contents [show]

Education [edit]

Hinton was educated at King's College, Cambridge graduating in 1970, with a Bachelor of Arts in experimen continued his study at the University of Edinburgh where he was awarded a PhD in artificial intelligence in 1

Inverse Cloze Task(ICT)

After his PhD he worked at the University of Sussex, and (after difficulty finding funding in Britain)^[26] the Universit San Diego, and Carnegie Mellon University.^[1] He was the founding director of the Gatsby Charitable Foundation Neuroscience Unit at University College London,^[1] and is currently^[27] a professor in the computer science depart University of Toronto. He holds a Canada Research Chair in Machine Learning, and is currently an advisor for the Machines & Brains program at the Canadian Institute for Advanced Research. Hinton taught a free online course Networks on the education platform Coursera in 2012.^[28] Hinton joined Google in March 2013 when his compan Inc., was acquired. He is planning to "divide his time between his university research and his work at Google".^[29] Hinton's research investigates ways of using neural networks for machine learning, memory, perception and sym He has authored or co-authored over 200 peer reviewed publications.^{[2][30]}

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to p specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training c one without being explicitly programmed to perform the task.[1][2]:2 Machine learning alc vision, where it is difficult or infeasible Wiki Link Prediction(WLP)

cuses on making predictions using compu of mathematical optimization delivers methods, theory and application domains to the field of machine learning. D field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning.^[3] application across business problems, machine learning is also referred to as predictive analytics.

Contents [show]

verview [edit]

e name machine learning was coined in 1959 by Arthur Samuel.^[5] Tom M. Mitchell provided a widely quoted, m definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experie respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, experience E."[6] This definition of the tasks in which machine learning is concerned offers a fundamentally operation rather than defining the field in cognitive terms. This follows Alan Turing's proposal in his paper "Computing Mach Intelligence", in which the question "Can machines think?" is replaced with the question "Can machines do what v entities) can do?".[7] In Turing's proposal the various characteristics that could be possessed by a thinking machin various implications in constructing one are exposed.

Machine learning tasks [edit]

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining contained a certain object, the training data for a supervised learning algorithm would include images with and wit input), and each image would have a label (the output) designating whether it contained the object. In special cas

- Natural language questions-answer passages vs. short queries-long documents
- Depend on document structure, e.g., WLP
- Marginal benefit for ad-hoc retrieval





Design a novel pre-training objective tailored for IR, which more closely resembles the relevance relationship between query and document.

Back to Statistical LM for IR

Classical SLM for IR: the Query Likelihood model



- •The user has a reasonable idea of the terms that are likely to appear in the "ideal" **document** that can satisfy his/her information need
- •The query is generated as the piece of text representative of the "ideal" document

Back to Statistical LM for IR

Query likelihood scoring function derived by the Bayesian theorem

$$P(D/Q) \propto P(Q)$$

Query generation probability Uniform distribution

- Smoothing methods for zero probability problem
 - E.g., Jelinek-Mercer, Dirichlet prior, Absolute discounting
 - Query likelihood with Dirichlet smoothing is one of the most effective method (Zhai et.al. 2001

$$P(\mathbf{q_i}|\theta_D) = \frac{c(w, D)}{|D|} \implies \frac{c(w, D) + \mu P(w|C)}{|D| + \mu}, \ \mu \text{ is small small small states}$$

 $(\theta_D)P(D) \propto P(Q|\theta_D)$ Document language model

oothing parameter, P(w|C) is collection language model



Pre-training Task for Ad-hoc Retrieval: ROP

- Representative words prediction (ROP) task

 - document
 - Pre-train the Transformer model to predict the representativeness

Overview [edit]

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a content collection or database. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching.^[1]

Depending on the application the data objects may be, for example, text documents, images,^[2] audio,^[3] mind maps^[4] or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the **query**.^[5]

From https://en.wikipedia.org/wiki/Information_retrieval

 Given a document, sample word sets according to the document language model • The word set with higher likelihood is deemed as more "representative" of the



Representative Word Sets Sampling

- 1. Given document *d*, initialize document language model with Dirichlet smoothing θ_d
- 2. Choose length $l \sim Poisson(\lambda)$
- **Paired Sampling**: Sample N pairs of word sets 3. for each document where $w_i \sim P(w_i | \theta_d)$
 - Why? Likelihood comparable
- Higher likelihood deemed as more 4. representative

Algorithm 1 Sampling a Pair of Representative Word Sets

- 1: Input:Document D, Vocabulary $V = \{w_i\}_{i=1}^{N}$, probability of word w_i generated by the document language model with Dirichlet smoothing $P(w_i|D)$, Query likelihood score $QL(w_i, D)$
- 2: //Sample length
- 3: $l = Sample(X), x \sim Poisson(\lambda), x = 1, 2, 3...$

$$4: S_1, S_2 = \emptyset, \emptyset$$

5: //Sample a pair of word sets according to the document language model

6: for $k \leftarrow 1$ to l do $S_1 = S_1 \cup Sample(V), w_i \sim P(w_i|D)$ 7: $S_2 = S_2 \cup Sample(V), w_i \sim P(w_i|D)$ 9: end for //Decide which one is more representative 11: $S_1_score = \prod_{i=1}^{l} QL(w_i, D), w_i \in S_1$ 12: $S_2_score = \prod_{i}^{l} QL(w_i, D), w_i \in S_2$ 13: if $S_1_score > S_2_score$ then **Output:** (S_1^+, S_2^-, D) 15: else **Output:** (S_1^-, S_2^+, D) 16: 17: end if

Pre-training with the ROP task

Pre-training Loss function



$$\mathcal{L}_{ROP} = \max(0, 1 - P(S_1|D) + P(S_2|D))$$
$$\mathcal{L}_{MLM} = -\sum_{\hat{x} \in X} \log p(\hat{x}|X_{\setminus \hat{x}})$$

Discussions

the labels are produced by some automatic model rather than simple MASKs.

- **Electra** leverages a generative model to replace masked tokens
- **PEGASUS** leverages the ROUGE1-F1 score to select top-m sentences \bullet

.



The ROP objective belongs to the category of model-based pre-training objective where

Experiment Setting

- Pretraining datasets:
 - Wikipedia, over 10 million documents
 - MS MARCO, about 3.4 million documents
- 5 downstream ad-hoc retrieval tasks: • Robust04, ClueWeb09-B, Gov2, MQ2007, MQ2008
- Baseline models:
 - Traditional retrieval models: BM25, QL
 - Previous state-of-the-art neural ranking models on each dataset: BERT-MaxP, HiNT et.al.
 - Other pretraining method: BERT, Transformer_{ICT}

Experiments – Main Results

Table 2: Comparisons between PROP and the baselines. *, \dagger and \ddagger indicate statistically significance with $p - value \le 0.05$ over **BM25**, **BERT** and **Transformer**_{*ICT*}, **respectively**.

Model	Robust04		ClueWeb09-B		Gov2		MQ2007		MQ2008	
	nDCG@20	P@20	nDCG@20	P@20	nDCG@20	P@20	nDCG@10	P@10	nDCG@10	P@10
QL	0.413	0.367	0.225	0.326	0.409	0.510	0.423	0.371	0.223	0.241
BM25	0.412	0.363	0.230	0.334	0.421	0.523	0.414	0.366	0.220	0.245
Previous SOTA	0.538	0.467	0.296	-	0.422	0.524	0.490	0.418	0.244	0.255
BERT	0.459*	0.389*	0.295*	0.367*	0.495*	0.586*	0.506*	0.419*	0.247*	0.256*
Transformer _{ICT}	0.460^{*}	0.388 *	0.298*	0.369*	0.499*†	0.587*	0.508*	0.420*	0.245^{*}	0.256*
PROP _{Wikipedia}	0.502 *†‡	0.421*†‡	0.316*†‡	0.384*†‡	0.519* ^{†‡}	0.593* ^{†‡}	0.523 * ^{†‡}	0.432 * ^{†‡}	0.262*†‡	0.267* ^{†‡}
PROP _{MSMARCO}	0.484* ^{†‡}	0.408*†‡	0.329 *†‡	0.391 *†‡	0.525 *†‡	0.594 *†‡	0.522*†‡	0.430*†‡	0.266 * ^{†‡}	0.269 *†‡

- 1. NDCG@20), except for the Robust04 (BERT + Neu-IR ensemble).
- PROP is significantly better than BERT and Transformer_{ICT}. 2.
- Pre-training in related domain corpus is more effective. 3.

PROP significantly outperforms previous SOTA in 4 of 5 tasks (8.9%, 24.4%, 6.7% and 9% in terms of



Experiments – Impact of Pre-training Objectives

Table 3: Impact of pre-training objectives. † indicates statis**tically significance with** *p* – *value* < 0.05.

		nDCG@20	nDCG@10		
	Robust04	ClueWeb09-B	Gov2	MQ2007	MQ2008
w/ MLM	0.467	0.306	0.503	0.511	0.249
w/ ROP	0.481 [†]	0.321 [†]	0.519†	0.520^{\dagger}	0.262^{\dagger}
w/ ROP+MLM	0.484 [†]	0.329 [†]	0.525 [†]	0.522 [†]	0.266 [†]

- 1.
- MLM and ROP are both helpful for downstream tasks. 2.

Pretraining with ROP achieves significant improvements over MLM.

Experiments – Impact of Sampling Strategies

docLM—based vs. Random sampling

Table 5: Impact of Different Sampling Strategies. Two-tailed t-tests demonstrate the improvements of document language model-based sampling to the random sampling strategy are statistically significant († indicates p-value < 0.05).

		nDCG@			
	Robust04	ClueWeb09-B	Gov2	MQ2007	M
Random	0.471	0.304	0.505	0.513	0
docLM-based	0.493 [†]	0.317^{\dagger}	0.517^{\dagger}	0.516^\dagger	0.

- docLM-based sampling converges faster and leads to better performance. 1.
- 2. generate representative word sets for a document.





Figure 1: (a) ROP learning curve on Wikipedia over the pretraining steps. (b) The test performance curve on Robust04 in terms of nDCG@20 over pre-training steps.

docLM-based sampling strategy is a more suitable way than the random sampling strategy to



21

Experiments – Zero-shot and few-shot setting



Figure 2: Fine-tuning with limited supervised data. The solid lines are PROP fine-tuned using 0 (zero shot), 10, 30, 50, and 70 queries for Robust04, ClueWeb09-B and Gov2 datasets, using 0 (zero shot), 50, 100, 150, and 200 queries for MQ2007 and MQ2008 datasets. The dashed lines are BERT fine-tuned using the full queries.

- 1. fine-tuned on the full supervised datasets, e.g., 30 queries on Robust04.
- Under the zero-shot setting, PROP also achieves exciting performance 2.
 - full dataset

PROP fine-tuned on limited supervised data can achieve comparable performance with BERT

On Gov2, PROP beats BM25 in terms of nDCG@20, and achieves about 90% performance of BERT fine-tuned on the

Conclusion & Future Work

- Conclusion

 - We proposed PROP, a new pre-training method tailed for ad-hoc retrieval • PROP achieved significant improvements over the baselines without pre-training or with other pre-training methods
 - PROP can achieve strong performance under both the zero- and low-resource IR settings
- Future work
 - Go beyond the ad-hoc retrieval, and test the ability of PROP over other downstream IR tasks, such as passage retrieval in QA or response retrieval in dialog systems • Investigate new ways to further enhance the pre-training objective tailored for IR



https://github.com/Albert-Ma/PROP

Thanks!

Xinyu Ma **Maxinyu17g@ict.ac.cn**

Code and the pre-training models are available at: