Unsupervised Document Expansion for Information Retrieval with Stochastic Text Generation

Soyeong Jeong[1], Jinheon Baek[2], ChaeHun Park[1], Jong C. Park[1*]

School of Computing[1], Graduate School of AI[2], KAIST[1,2], South Korea

{syjeong, ddehun, park}@nlp.kaist.ac.kr, jinheon.baek@kaist.ac.kr

Motivation

One of the challenges in information retrieval (IR) is the *vocabulary mismatch problem*, which happens when the terms between queries and documents are lexically different but semantically similar. While recent work has proposed to expand the queries or documents by enriching their representations with additional relevant terms or densely represent documents to address this challenge, they usually require a large volume of query-document pairs to train an expansion model.

UDEG Framework

To tackle the vocabulary mismatch problem, we propose an Unsupervised Document Expansion with Generation (UDEG) framework with a pre-trained language model, which generates diverse supplementary sentences for the original document without using labels on query-document pairs for training. We first generate document-related sentences with a pre-trained language model, already fine-tuned on a summarization task. However, such a framework generates only one static sentence at a time, so we further propose to stochastically generate multiple sentences which reflect diverse points of view for the given document and minimize the vocabulary mismatch cases (See Figure 1).

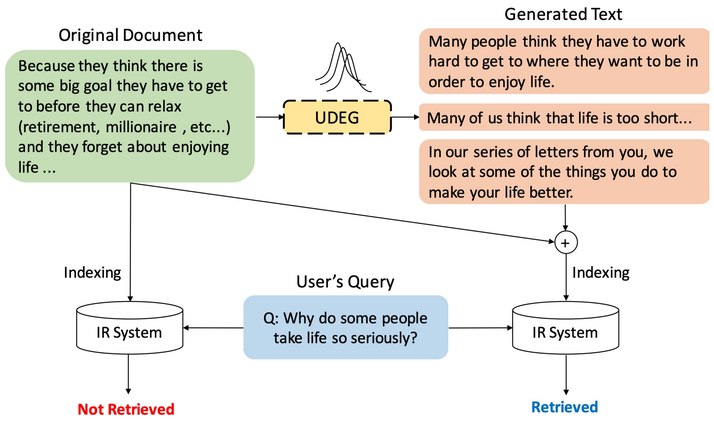

Figure 1. UDEG framework, where the example is generated from UDEG. Given an original document (green box), our UDEG framework stochastically generates several sentences (orange box) relevant to the given document, and augments the generated sentences to the input document.

Baseline Expansion Models

- No Expansion
- Query Expansion Model
  - RM3
- Extractive Document Expansion Models
  - MP-rank
  - LexRank
  - Pegasus$_{ext}$
- Abstractive Document Expansion Models
  - LexRank + paraphrase
  - UDEG

Main Results

UDEG outperforms other baselines on the ANTIQUE dataset and sampled MS MARCO dataset (See Table 1 and Table 2).

		No Expan.	MP-rank	LexRank	Lex.+Para.	Pegasus$_{ext}$	UDEG
MRR	BM25	0.595	0.584	0.571	0.561	0.585	**0.645**
	BM25+RM3	0.558	0.579	0.542	0.567	0.555	**0.616**
	QL	0.499	0.534	0.567	0.518	0.562	**0.650**
	QL+RM3	0.396	0.447	0.456	0.432	0.504	**0.583**
R@10	BM25	0.218	0.220	0.208	0.209	0.207	**0.237**
	BM25+RM3	0.217	0.221	0.208	0.204	0.213	**0.226**
	QL	0.189	0.199	0.203	0.196	0.205	**0.232**
	QL+RM3	0.159	0.179	0.182	0.162	0.191	**0.211**
P@3	BM25	0.378	0.381	0.346	0.351	0.356	**0.431**
	BM25+RM3	0.361	0.355	0.360	0.373	0.366	**0.433**
	QL	0.301	0.333	0.340	0.315	0.358	**0.418**
	QL+RM3	0.240	0.281	0.275	0.271	0.301	**0.386**
MAP	BM25	0.211	0.212	0.199	0.202	0.201	**0.238**
	BM25+RM3	0.212	0.213	0.203	0.203	0.207	**0.234**
	QL	0.172	0.191	0.192	0.181	0.199	**0.230**
	QL+RM3	0.150	0.168	0.170	0.158	0.180	**0.212**
NDCG@3	BM25	0.437	0.442	0.417	0.425	0.419	**0.478**
	BM25+RM3	0.424	0.434	0.423	0.433	0.426	**0.470**
	QL	0.356	0.389	0.400	0.375	0.413	**0.471**
	QL+RM3	0.277	0.324	0.319	0.306	0.350	**0.424**

Table 1. Retrieval results on the ANTIQUE dataset.

		No Expan.	LexRank	UDEG
MRR	BM25	0.427	0.441	**0.463**
	BM25+RM3	0.366	0.385	**0.415**
	QL	0.402	0.420	**0.454**
	QL+RM3	0.319	0.337	**0.382**
R@10	BM25	0.636	0.646	**0.679**
	BM25+RM3	0.600	0.617	**0.651**
	QL	0.611	0.633	**0.671**
	QL+RM3	0.552	0.579	**0.629**
P@1	BM25	0.311	0.324	**0.344**
	BM25+RM3	0.248	0.265	**0.291**
	QL	0.289	0.302	**0.334**
	QL+RM3	0.202	0.215	**0.255**
MAP	BM25	0.422	0.435	**0.457**
	BM25+RM3	0.361	0.380	**0.409**
	QL	0.398	0.414	**0.448**
	QL+RM3	0.315	0.333	**0.377**

Table 2. Retrieval results on the MS MARCO dataset.

Ablation Study

- Robustness on different LMs (See Figure 2).
- Comparison of Stochastic Generation Strategy (See Figure 3).
- Varying the Number of Expanded Sentences (See Figure 3).

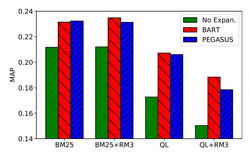

Figure 2. Comparison of BART and PEGASUS.

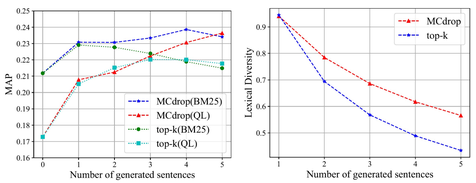

Figure 3. MC dropout vs top-k sampling (Left) with a varying number of generated sentences and (Right) with a lexical diversity.

Case Study

UDEG successfully retrieves documents (See Table 3).

Query	How is the chemistry is a basic of science?
Relevant Document	Chemistry is a basic because all matter can be broken down into elements (i.e., hydrogen, oxygen, nitrogen, etc.); without matter, nothing could be studied.
Generated Sentences	**1)** Chemistry is the study of atoms and molecules. **2)** Chemistry is the study of matter and how it is made. **3)** Chemistry is the study of matter. **4)** Chemistry is a basic **science**.
	Original Document Rank: 104 Expanded Document Rank: 5
Query	How is the library consider as a heart of university?
Relevant Document	Whatever you are studying has to be found somewhere for you to learn it. That's where the library comes into focus.
Generated Sentences	**1)** If you're studying at **university**, you'll need a library. **2)** A library is a place where you can find out more about the subject you are studying. **3)** If you're studying, you'll be studying. **4)** There are many different ways you can study.
	Original Document Rank: 636 Expanded Document Rank: 32
Query	What do doctors do when a patient has a Do Not Resuscitate Order?
Relevant Document	All healthcare professionals involved in the care of that patient will not do anything to prolong the patient's life if in case patient deteriorates/dies. DNR orders may be modified, some may choose mechanical ventilation, or drugs. Usually when a pt is DNR, comfort measures is provided only.
Generated Sentences	**1)** DNR is not life-support. **2)** When a patient is in a "do not resuscitated" (DNR) state, that patient's life will not be saved. **3)** A DNR is a decision made by the patient's family or health care provider to prolong the life of the patient. **4)** A "do not resuscitate"(DNR) order does not mean that a patient should be put on life support.
	Original Document Rank: 40 Expanded Document Rank: 1

Table 3. Examples of generated sentences on the ANTIQUE dataset.

Contributions

- To mitigate the vocabulary mismatch problem, we present UDEG framework that augments a document with abstractly generated sentences without paired query-document data.
- Under an unsupervised document expansion framework, we generate document-related sentences with a pre-trained LM, and further stochastically generate diverse sentences.
- We show that our framework achieves outstanding performances on benchmark datasets for IR tasks.