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).

### Baseline Expansion Models

- No Expansion
- Query Expansion Model
- RM3
- Extractive Document Expansion Models
  - MP-rank
  - LexRank
  - Pegasus
- 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).

#### Case Study

UDEG successfully retrieves documents (See Table 3).

#### 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).

### 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.