From Fundamentals to Recent Advances A Tutorial on Keyphrasification

Part 3.1 Keyphrase Generation for Information Retrieval

Rui Meng, Debanjan Mahata, Florian Boudin

ECIR 2022

salesforce research

Moody's Analytics

LS2N Laboratoire des Sciences du Numérique de Nantes

Keyphrase Generation for IR

- Keyphrases distill the **most important information** from documents
- So, what about using keyphrases to improve IR effectiveness?
- Use keyphrases to supplement document indexing
  - Adding keyphrases = **document expansion**
  - 👍 Improves document retrievability (addresses the vocabulary mismatch problem)
  - 👎 Larger index & possible document drift
- **Extrinsic evaluation framework** for keyphrase generation/extraction

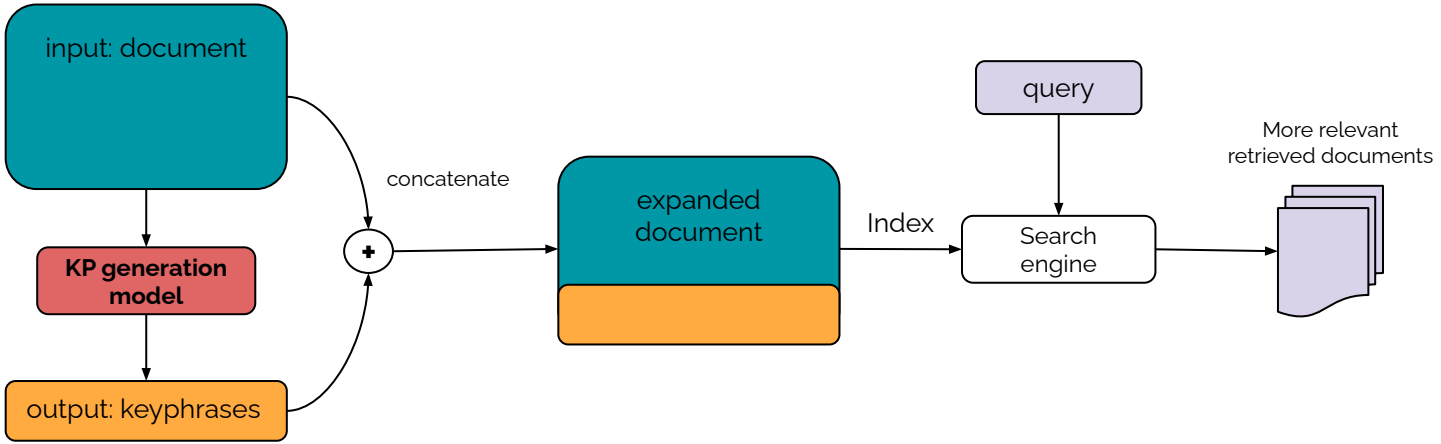
Keyphrase Generation for IR (cont.)

🤔 Why bothering ? we have neural IR models, i.e. semantic level matching

- Scalable neural IR models use exact term match to perform initial retrieval
- In a zero-shot setup
  - neural IR models underperform lexical models (Thakur et al., 2021)
  - neural document expansion shows consistent improvements (Nogueira and Lin, 2019)
- Interpretability and error-correcting

(Thakur et al., 2021) N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS 2021.
(Nogueira and Lin, 2019). R. Nogueira and J. Lin. From doc2query to docTTTTTquery. 2019.

Keyphrase Generation for IR (cont.)



Results

- Experiments conducted on the NTCIR-2 test collection
  - Ad-hoc retrieval models implemented in the Anserini toolkit
  - Two neural kp generation methods + unsupervised extractive baseline

Index	BM25	+RM3
Title+Abstract	29.16	31.93
+ s2s+copy	30.54†	**34.30**†
+ s2s+corr	30.30†	33.24
+ mp-rank	29.24	32.27
Title+Abstract+Keywords	31.38	35.17
+ s2s+copy	31.55	**36.53**‡
+ s2s+corr	31.37	35.84
+ mp-rank	31.38	35.18

Findings from (Boudin et al., 2020)

1. using kps improves retrieval effectiveness
2. gains when using both generated + author kps
3. gains of query and document expansion are additive

(Boudin, 2020) F. Boudin, Y. Gallina, A. Aizawa. Keyphrase generation for scientific document retrieval. ACL 2020..

A closer look at present/absent keyphrases for IR

- Keyphrases **occur** or **not** in the document
  - ≈ **60% present** - **40% absent** in author kps
- Keyphrases have different effects on IR
  - **present kps** → term re-weighting
  - **absent kps** → document expansion

🤔 To which extent present/absent kps affect retrieval effectiveness?

Sample document (gakkai-e-0001384947) from NTCIR-2

Study on the Structure of Index Data for Metasearch System

This paper proposes a new technique for **Metasearch system**, which is based on the grouping of both keywords and URLs. This technique enables **metasearch** systems to share information and to reflect the estimation of users' preference. With this system, users can search not only by their own keywords but by similarity of HTML documents. In this paper, we describe the principle of the grouping technique as well as the summary of the existing **search systems**.

Keyphrases: Metasearch - Search System - Information Sharing - Information Retrieval - User's Behavior - Retrieval Support

What makes a keyphrase absent?

- Kps that do not match any contiguous subsequence of source text (Meng et al., 2017)
- From an IR perspective
  - stemmed words are used to index documents
  - **absent kps** can have **some** or **even all of their words occurring** in the source text
  - unseen words: [**retrieval, behavior, support**]
- We need to **re**define what are absent kps in the context of IR

Sample document (gakkai-e-0001384947) from NTCIR-2

Study on the Structure of Index Data for Metasearch System

This paper proposes a new technique for **Metasearch system**, which is based on the grouping of both keywords and URLs. This technique enables **metasearch** systems to **share information** and to reflect the estimation of **users'** preference. With this system, **users** can search not only by their own keywords but by similarity of HTML documents. In this paper, we describe the principle of the grouping technique as well as the summary of the existing **search systems**.

Keyphrases: Metasearch - Search System - Information Sharing - Information Retrieval - User's Behavior - Retrieval Support

(Meng et al., 2017) R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, Deep keyphrase generation, ACL.

The PRMU scheme for categorizing absent keyphrases

Present: kps that match contiguous sequences of words in the source document

Reordered: kps whose constituent words occur in the source document but not as contiguous sequences

Mixed: kps from which some, but not all, of their constituent words occur in the source document

Unseen: kps whose constituent words do not occur in the source document

Sample document (gakkai-e-0001384947) from NTCIR-2

Study on the Structure of Index Data for Metasearch System

This paper proposes a new technique for **Metasearch system**, which is based on the grouping of both keywords and URLs. This technique enables **metasearch** systems to **share information** and to reflect the estimation of **users'** preference. With this system, **users** can search not only by their own keywords but by similarity of HTML documents. In this paper, we describe the principle of the grouping technique as well as the summary of the existing **search systems**.

Present keyphrases: Metasearch - Search System

Absent keyphrases: Information Sharing - Information Retrieval - User's Behavior - Retrieval Support

Reordered *Mixed* *Mixed* *Unseen*

The proposed PRMU scheme draws a distinction between kps that **expand the document** (M+U) and **those that don't** (P+R)

(Meng et al., 2017) R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, Deep keyphrase generation, ACL.

Distribution of keyphrases under the PRMU scheme

	absent keyphrases				
Dataset	**%P**	**%R**	**%M**	**%U**	**%uw**
NTCIR-2	61.9	8.1	16.5	13.5	21.4
ACM-CR	53.6	11.7	19.3	15.4	25.5
KP20k	60.2	9.5	15.4	15.0	22.3
	term-weighting		doc. expansion		

ratio of unique, unseen words in kps

- Similar distributions across datasets, with absent kps ≈40%
- Most of the absent kps belong to M and U categories
- only ≈20% of the words included in kps contribute to expanding documents

Results

- Effect of indexing PRMU on retrieval effectiveness
  - Ad-hoc retrieval models implemented in the Anserini toolkit / NTCIR-2 test collection

index	BM25	+RM3	#kp
title & abstract	29.55	32.83	-
+ Present	30.74†	33.47	2.9
+ Reordered	29.79	33.48	0.4
+ Mixed	**30.80**†	33.85	0.8
+ Unseen	29.67	**33.94**	0.7
+ Absent (R+M+U)	30.77†	34.87†	1.9
+ Highlight (P+R)	30.64†	33.82	3.3
+ Expand (M+U)	**30.83**†	**34.34**	1.5
+ all (P+R+M+U)	31.92†‡	35.48†‡	4.8

Findings from (Boudin et al., 2021)

1. Largest gains with M & U
   - comparatively small number of kps
   - expanding document > highlighting phrases
2. Output distributions of keyphrase generation models are heavily skewed towards present kps
   - more work towards generating absent kps !

Further links

- To evaluate your keyphrase generation model through IR
  - https://github.com/boudinfl/ir-using-kg
- A manually curated test collection for citation recommendation
  - https://github.com/boudinfl/acm-cr
- The PRMU scheme for keyphrases
  - https://github.com/boudinfl/redefining-absent-keyphrases