Structure-Augmented Keyphrase Generation

Jihyuk Kim Myeongho Jeong Seungtaek Choi YONSEI UNIVERSITY Seung-won Hwang SEOUL NATIONAL UNIVERSITY

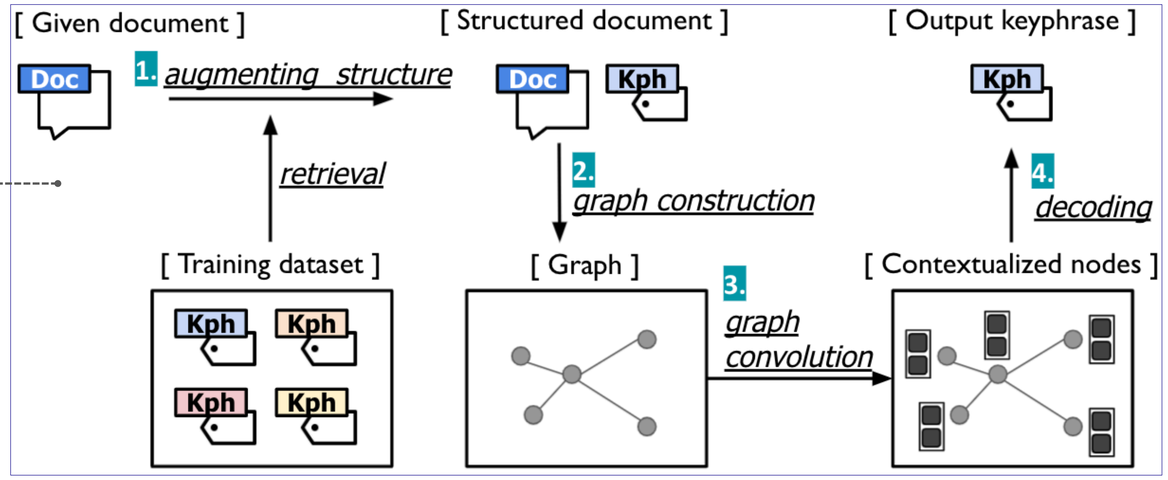
Summary

- We study **keyphrase generation from structured documents**.
- Our goal is to **augment/generate structures** for the given document, using **existing keyphrases**.
- We devise graphs that **effectively integrate** the given document and the retrieved keyphrases.

Task: Keyphrase Generation (KG)

Keyphrase: Keyphrases, also called hashtags, are short text segments that summarize the main contents of the given document.

Keyphrase Generation: KG aims to generate keyphrases from the given document.



Proposal: Structure-Augmented Keyphrase Generation

Previous work: title-body structures

Title

- Similar to keyphrases, titles summarize documents.

Challenges

- Challenge 1. Titles are short!
  - Titles may exclude some meaningful keywords.
- Challenge 2. Titles may not exist at all!
  - For some social media posts, e.g., Tweets, there is no titles.

Ours: leveraging existing keyphrases

Research Questions

- **How to complement incomplete titles.**
- **How to replace titles when those are not available.**

Leveraging existing keyphrases

- To augment/generate structures, we **leverage existing keyphrases** of other documents, which can be easily obtained.

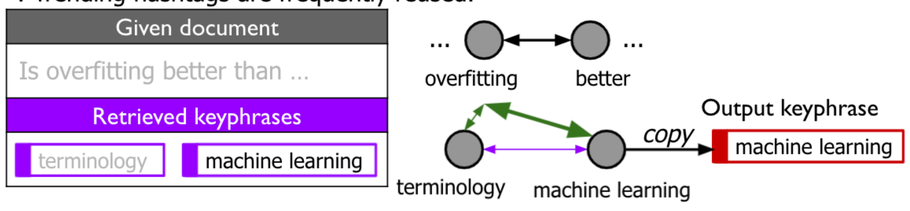
Graph construction for closed/open set keyphrases

Graph construction principle

- Relevant **contexts should be exchanged** between the given document and the retrieved keyphrases.

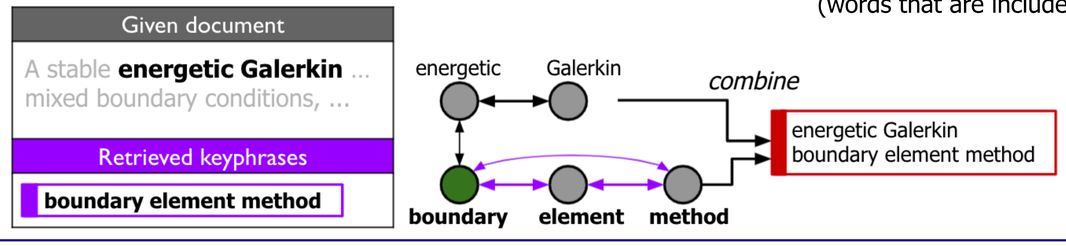
Closed set (e.g., social media posts)
: Trending hashtags are frequently reused.

Context exchange via **inter-field edges**



Open set (e.g., scientific publications)
: New terms are introduced continuously.

Context exchange via **merged nodes**
(words that are included in the both fields)



Experiment

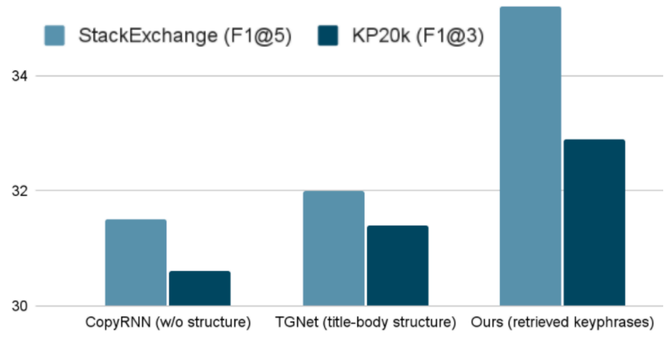
Datasets: StackExchange (social Q&A posts; closed set), **KP20k** (scientific publications; open set)

Baselines: **CopyRNN** and **TGNet** that use plain texts inputs and title-body structures respectively.

Evaluation metric: **F1 scores for top-k keyphrase predictions**.

Evaluation results

- CopyRNN < TGNet < **Ours**
  : Leveraging existing keyphrases to augment/generate structured documents are effective.



- Kph (ours) =< **Kph + Title (ours)**
  : Titles complement the retrieved keyphrases, when those are less relevant.