Web Document Encoding for Structure-Aware Keyphrase Extraction

Jihyuk Kim (YONSEI UNIVERSITY) Young-in Song (NAVER) Seung-won Hwang (SEOUL NATIONAL UNIVERSITY)

Summary

- We study **keyphrase extraction on structured Web documents**.
- Our goal is to leverage **complementary benefits of multiple fields** in structured Web documents.
- **Unified graph** with **intra-/inter-field edges** is employed to represent multi-field Web documents.

Proposal: Structure-Aware Encoding on Multi-Field Web Document

Keyphrases

- duty-free shopping
- pre-ordering discount
- pre-ordering benefits

Multi-Field Web Document

Fields	Text	Benefits
Title	duty-free shopping	**high precision** (low recall)
Header	duty-free | discount coupon	
Body	• all pre-ordering customers • 1 coupon per USD 400 • … • you can use the coupon on your next purchase of at least • …	**high recall** (low precision)
Anchor	pre-ordering purchase *benefits*	***additional vocabulary***

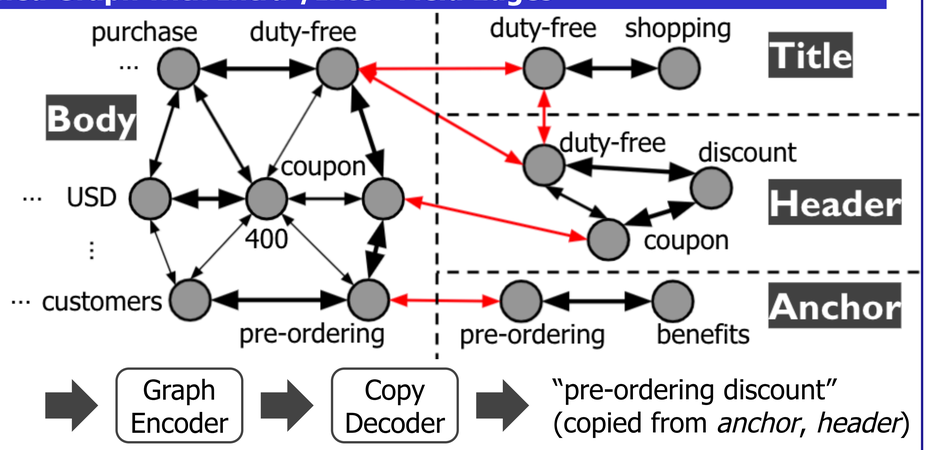
Research Questions

- **RQ1. How to model different language characteristics between fields.**
- **RQ2. How to enjoy complementary benefits between fields.**

Unified Graph with Intra-/Inter-Field Edges

Graph Construction

- **Nodes** are words in the given document
- **Intra-field edges** connect words within each field, with *position-based proximity* (thickness of intra-field edges).
- **Inter-field edges** connect words across fields, *having the same word*.



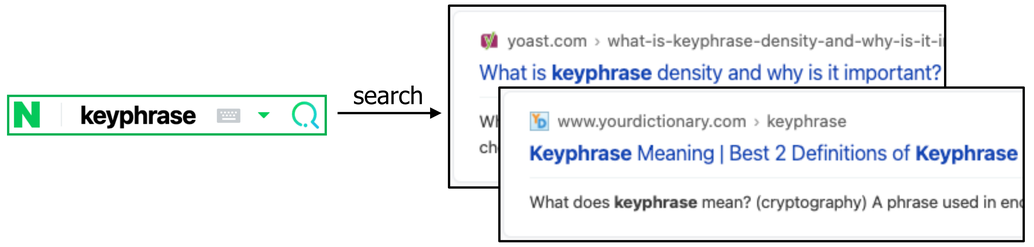
Graph Encoding using Graph Convolution Network (GCN)

- **RQ1.** To model different language characteristics, we use *different GCN parameters between fields*.
- **RQ2.** To model inter-field relations, we enable *the contexts to be exchanged across fields via inter-field edges*.

Task: Keyphrase Extraction (KE)

Keyphrase Extraction: KE extracts keyphrases from the given document.

Application: Keyphrases can improve **document ranking**, by emphasizing important contents.



Experiment: Document Ranking

Dataset

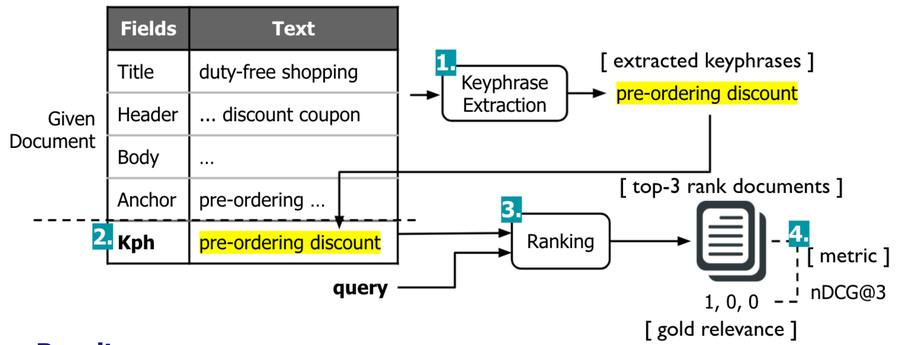
- **Click queries**, obtained using NAVER search engine, of a Web document are treated as keyphrases.

Compared Encoders

- **SeqEnc** uses a single layer GRU on Web documents without structures.
- **TGEnc** uses title-guided GRU using a simple title-body structure.
- **GraphEnc** uses GCN encoder on Web documents without structures.
- **MFGraphEnc (ours)** uses GCN encoder on multi-field Web documents.

Evaluation: Document Ranking

- We first extract a keyphrase from the given Web document, then use the extracted keyphrase, as an additional field of the document, to improve ranking.



Evaluation Results

- 1) **Doc w/o Kph** < **Doc w/ Kph**; Extracted keyphrases improve document ranking.
- 2) **MFGraphEnc** is the best; Leveraging structures further improves performance.