Web Document Encoding for Structure-Aware Keyphrase Extraction

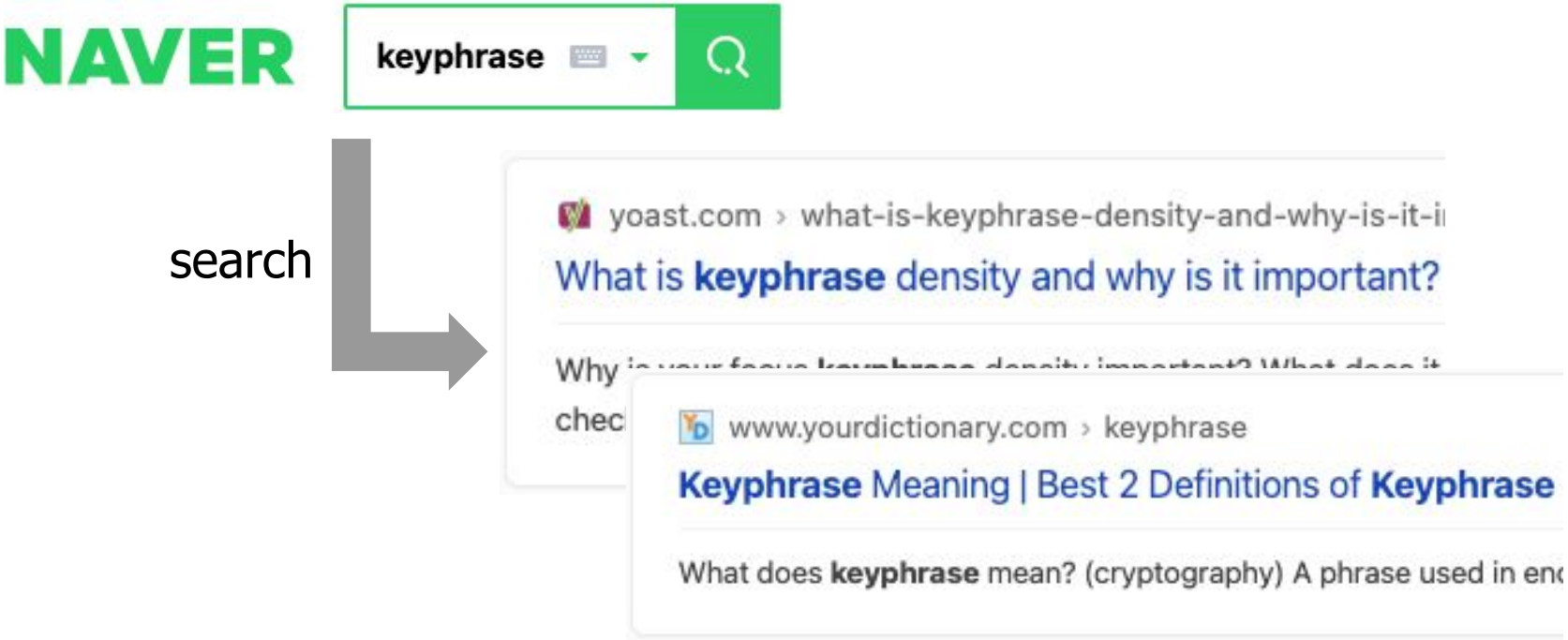
Jihyuk Kim
Yonsei Univ

Young-In Song
Naver Corp

Seung-won Hwang
Seoul National Univ

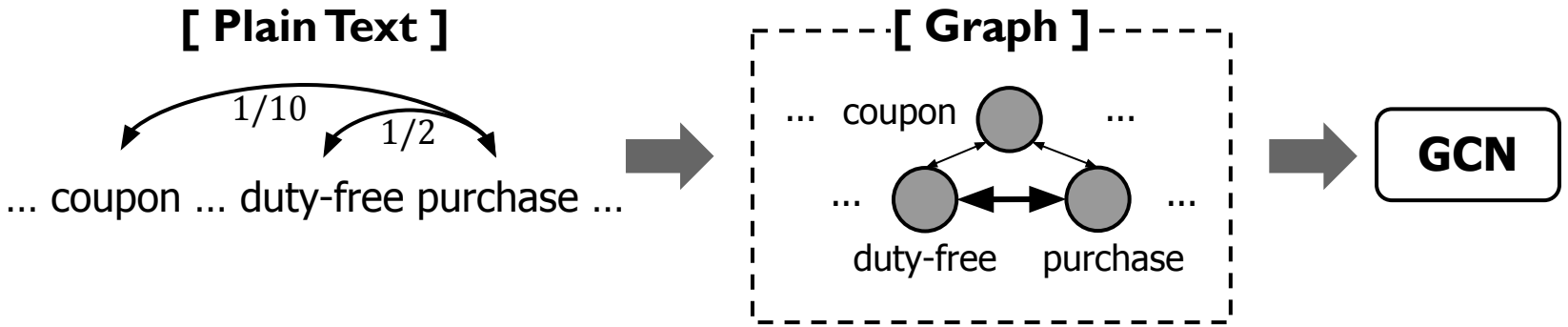
Task: Keyphrase Extraction for Web Document

- We aim to **extract keyphrases** that describe the main contents of a document.
  - Extracted keyphrases can improve **document ranking**.



Previous Work: GCN for Plain Text Encoding

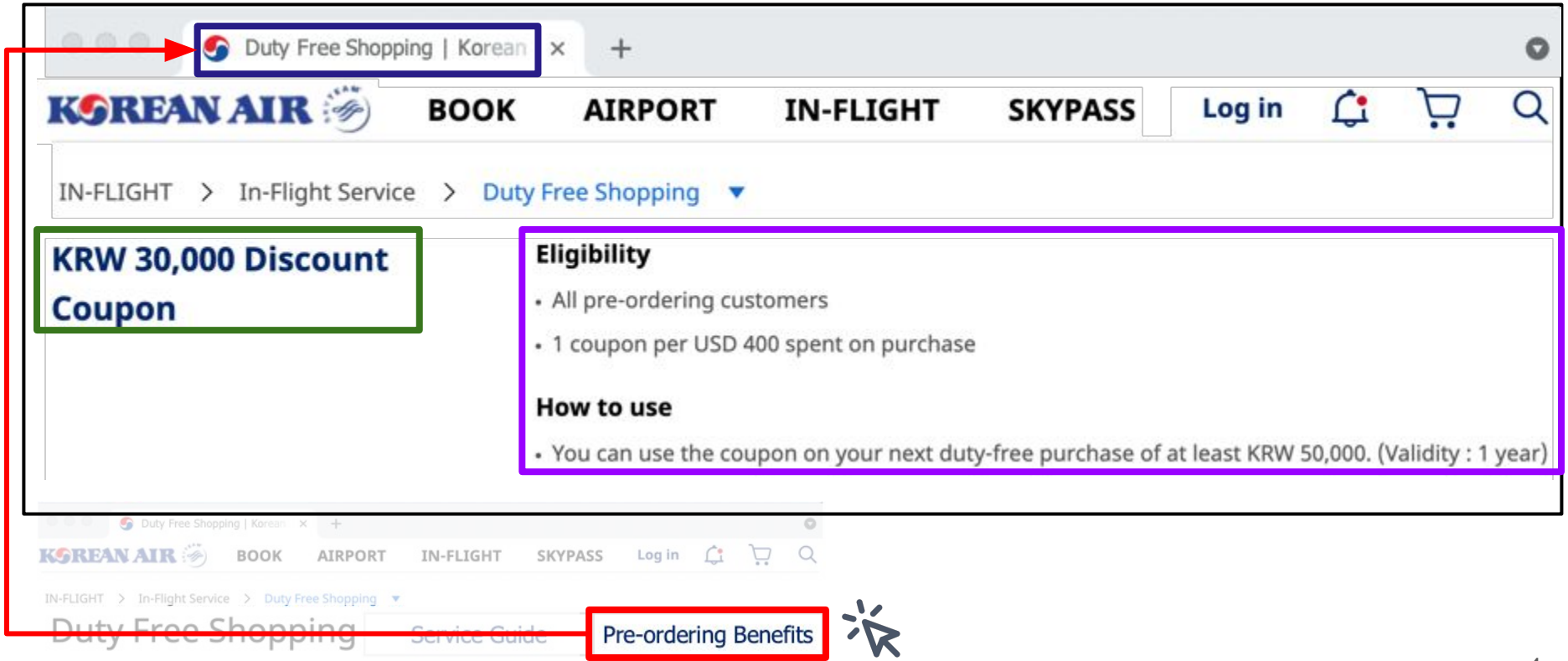
- DivGraphPointer[1] use graph representation for encoding a *plain text*.
  - **A fully connected graph** is constructed on word nodes.
  - For edge weights, **position-based proximities** are used.
  - **G**raph **C**onvolutional **N**etwork (**GCN**) is adopted to contextualize the graph.



[1] DivGraphPointer: A graph pointer network for extracting diverse keyphrases. Sun et al., SIGIR 2019

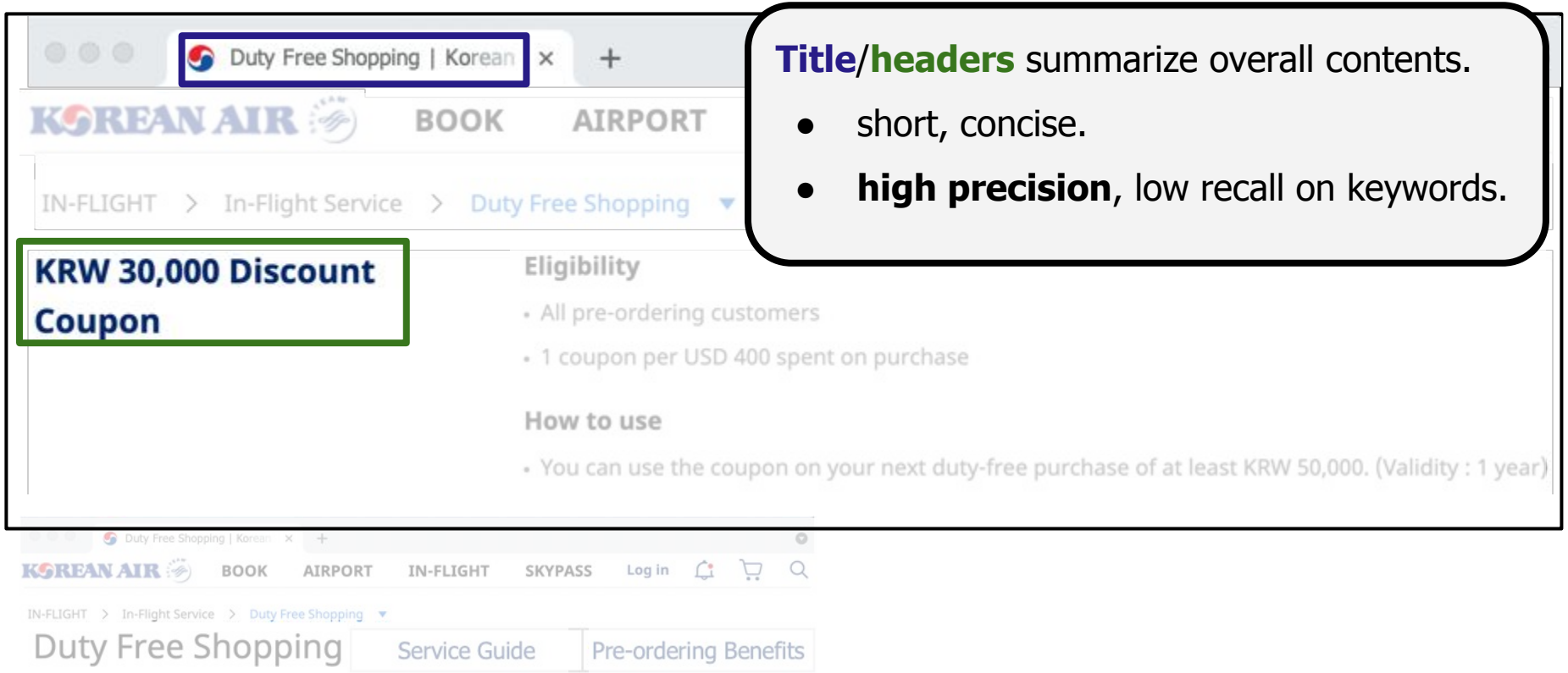
Motivation: Structure in Web Document, beyond Plain Text

- Web documents consist of **multiple fields**, e.g., title, header, body, or anchor text.



Motivation: Multiple Fields with Complementary Benefits

- Those fields provide **complementary benefits**.

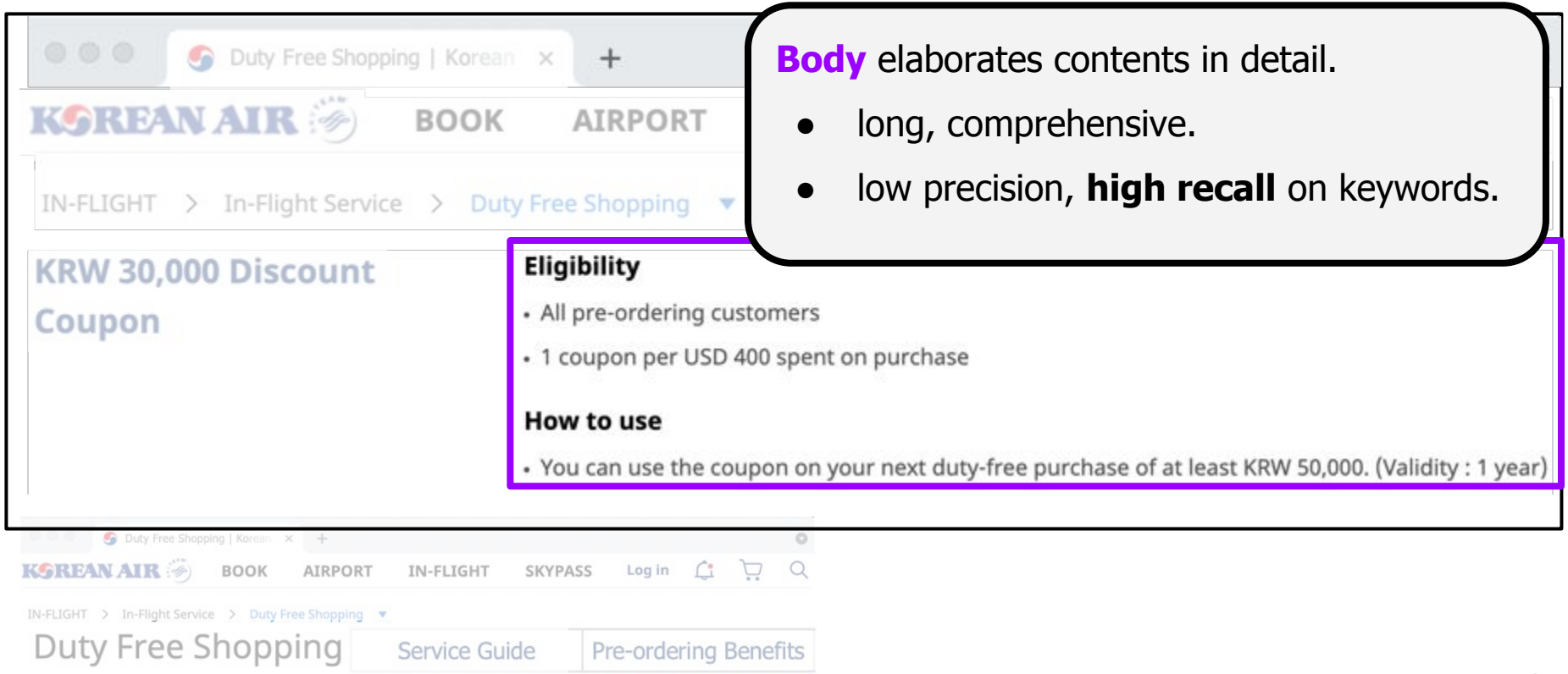


Title/**headers** summarize overall contents.

- short, concise.
- **high precision**, low recall on keywords.

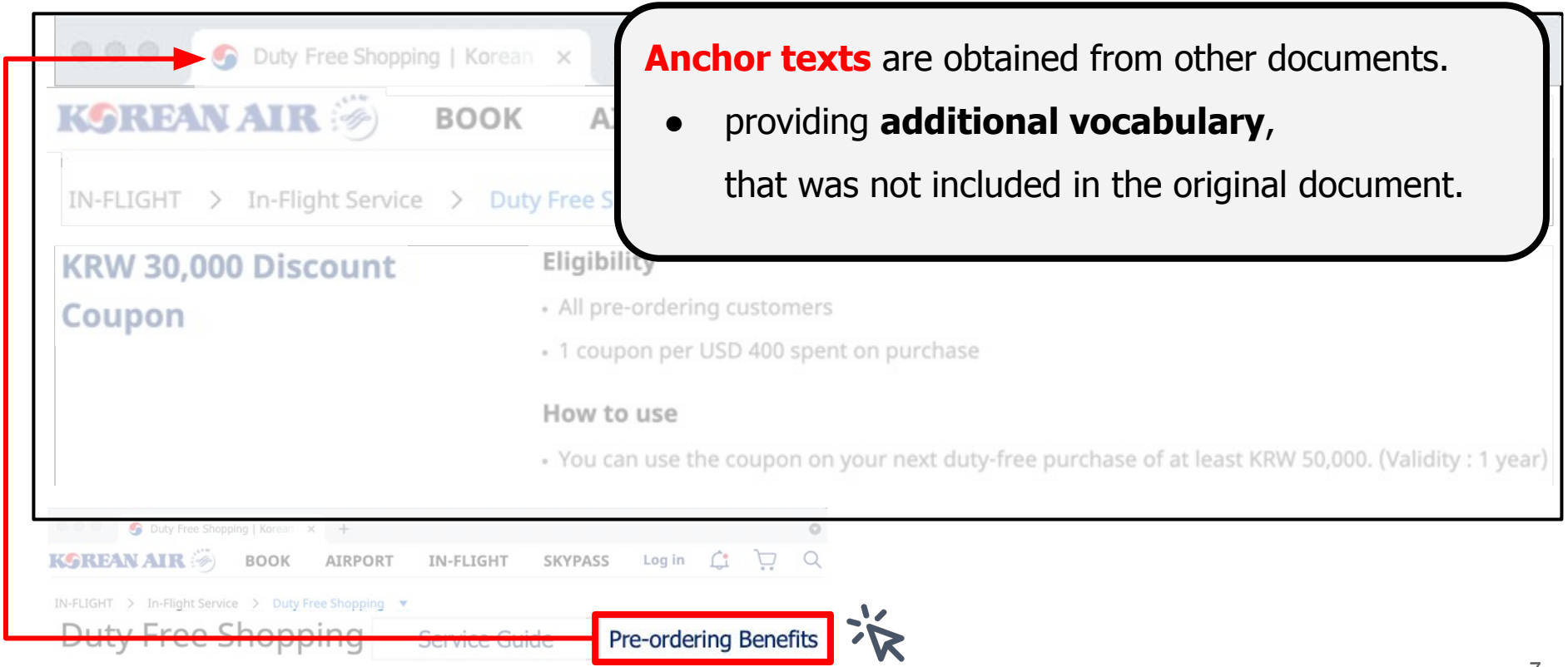
Motivation: Multiple Fields with Complementary Benefits

- Those fields provide **complementary benefits**.



Motivation: Multiple Fields with Complementary Benefits

- Those fields provide **complementary benefits**.



Anchor texts are obtained from other documents.

- providing **additional vocabulary**,
  that was not included in the original document.

Motivation: Multiple Fields with Complementary Benefits

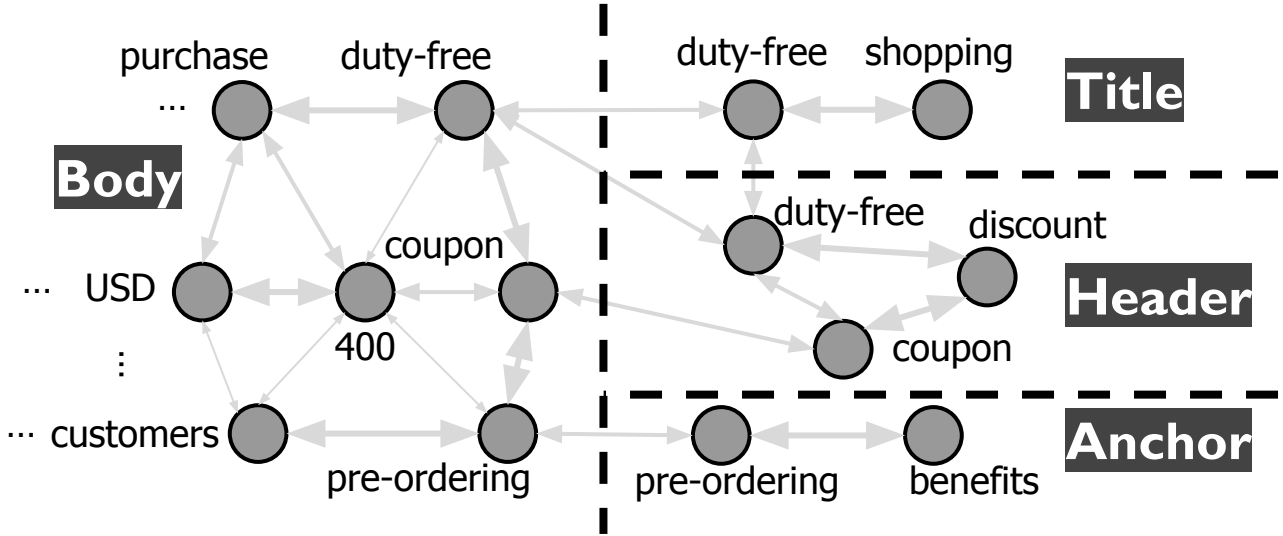
- Those fields provide **complementary benefits**.

Our goal is **structure-aware encoding**:

- 1) to **model different language characteristics between fields**,
- 2) and to **model inter-field relation**, to enjoy complementary benefits.

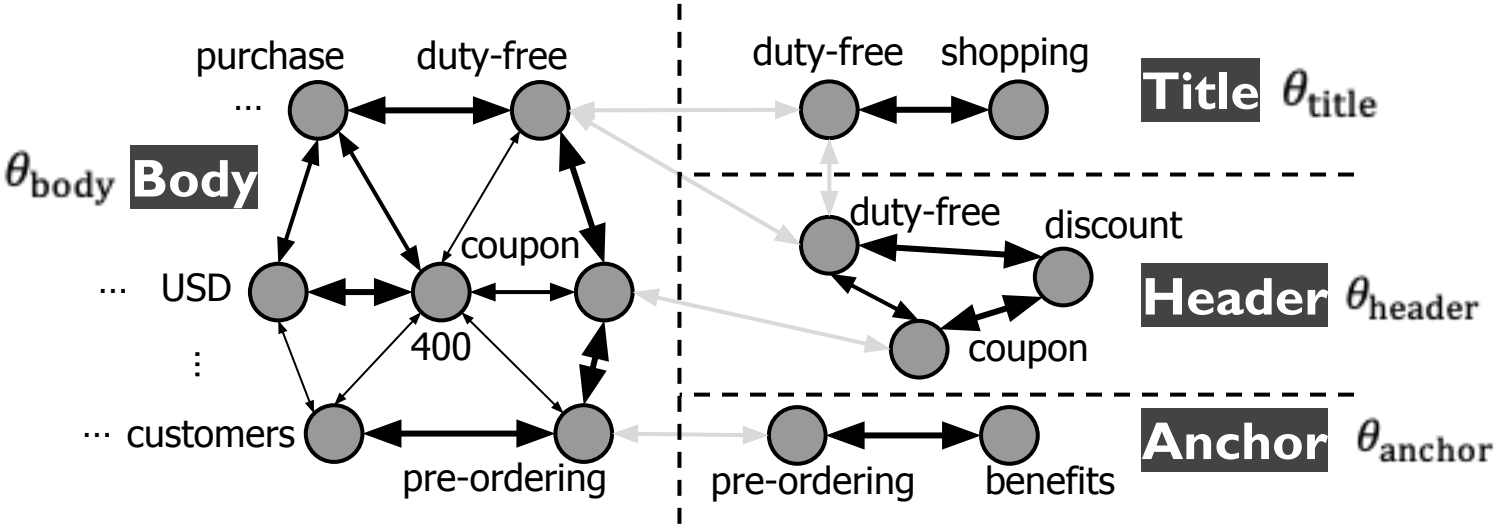
Approach: Multi-Field Graph Encoding

- We represent multiple field contents using an **unified graph**,
  - consisting of **multiple sub-graphs**, where each graph corresponds to each field.



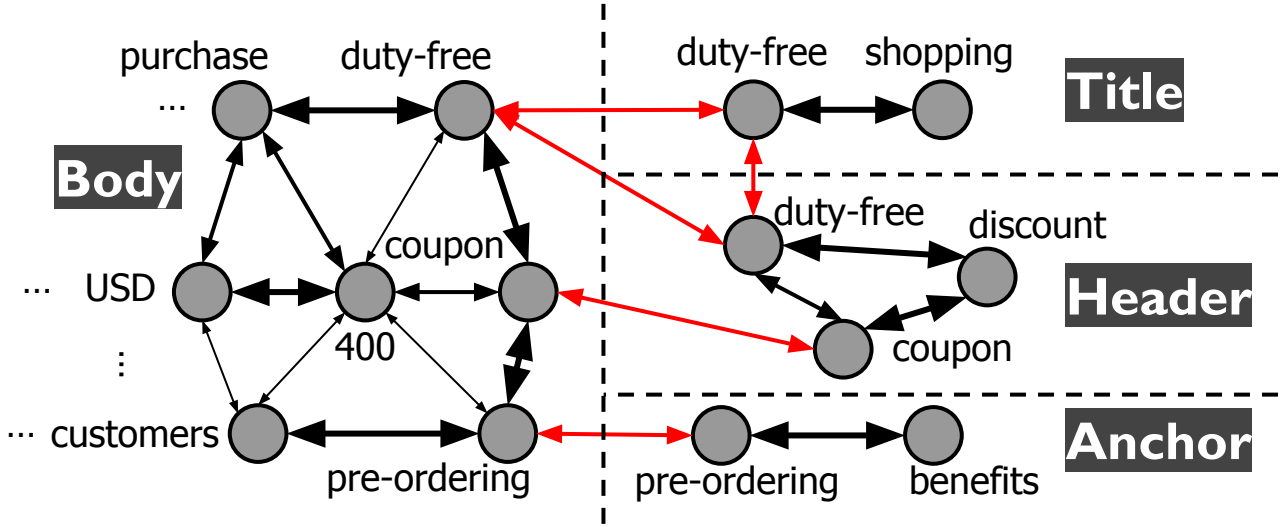
Approach: Multi-Field Graph Encoding

- We represent multiple field contents using an unified graph.
  - **Intra-field** edges (⟷) between words within each field.
    - We use **different GCN parameters** between fields.



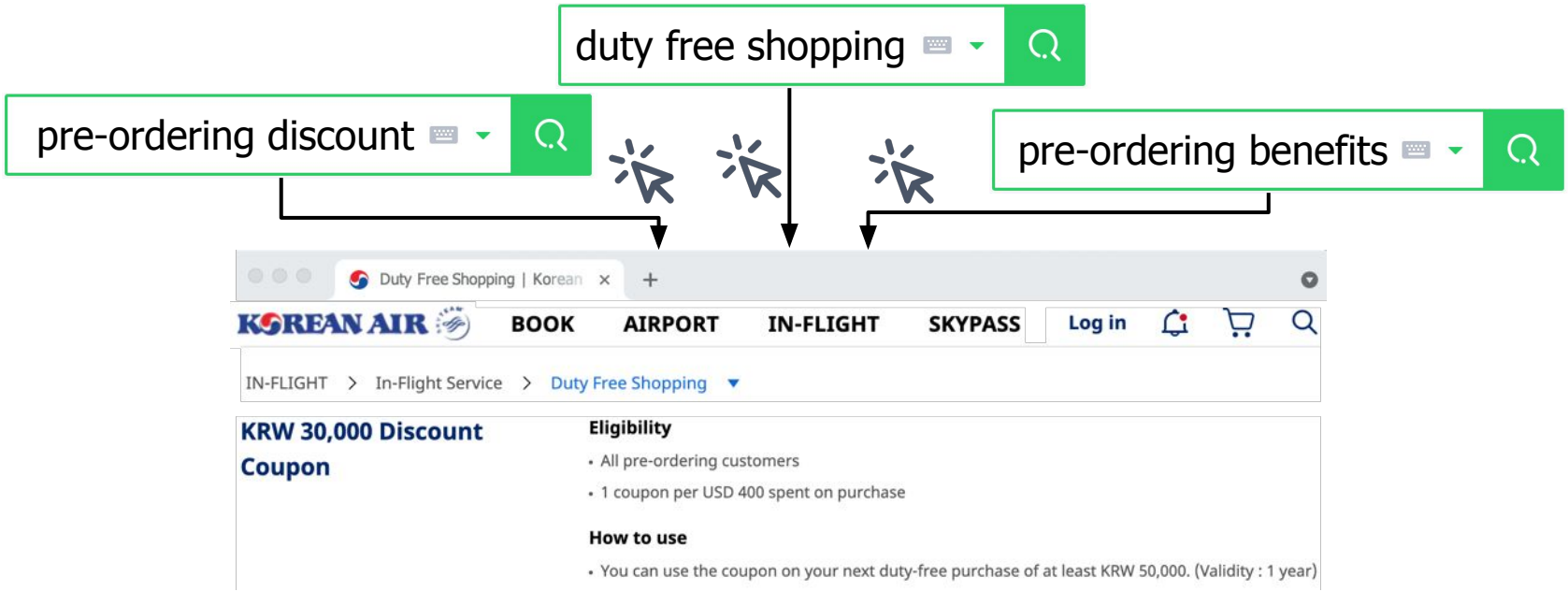
Approach: Multi-Field Graph Encoding

- We represent multiple field contents using an unified graph.
  - **Inter-field** edges (⟷) between words from different fields.
    - through which, **inter-field relations** can be modeled.



Experiment - Dataset

- We use real-world Web documents for experiments.
  - As keyphrases, we use "**click queries**" of Web documents.
    - click queries are gathered using NAVER search engine.



Experiment - Model Comparison

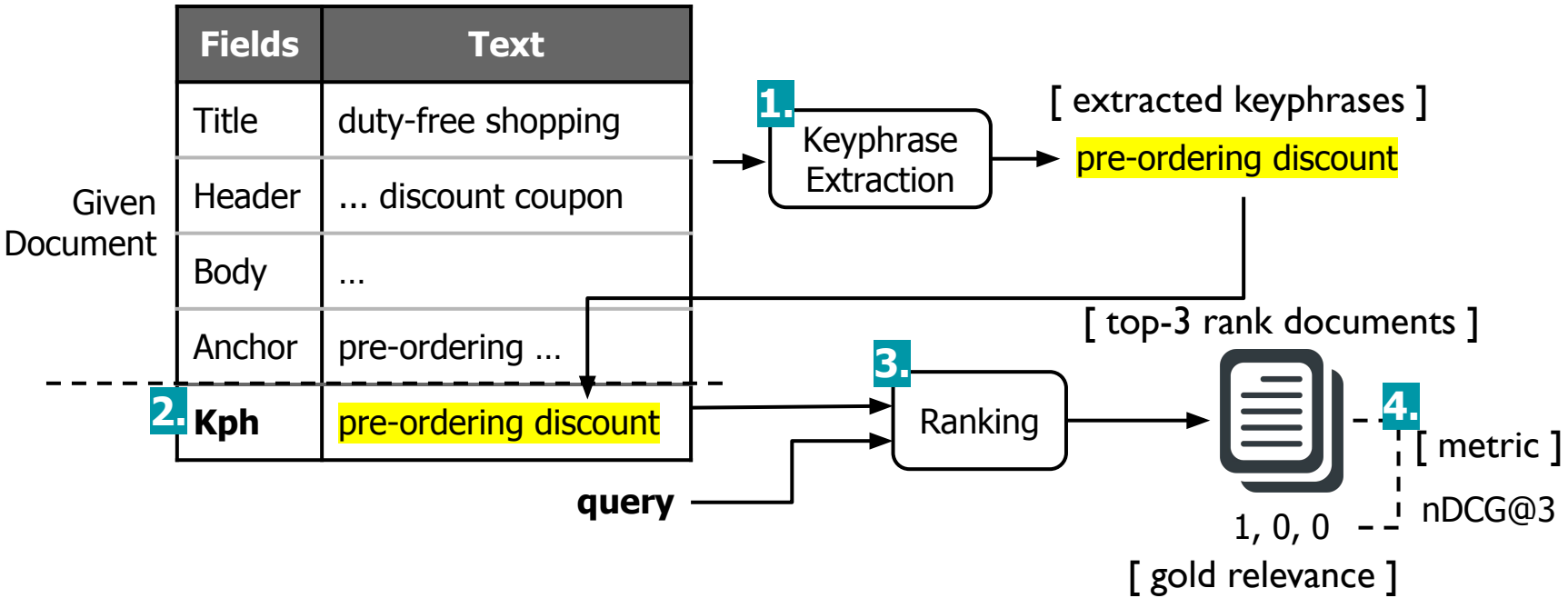
- We compare encoders with and without structures.

Model		Encoder
Baseline	GraphEnc[1]	GCN w/o structure
Ours	**MFGraphEnc**	**Multi-Field GCN w/ structure**

[1] DivGraphPointer: A graph pointer network for extracting diverse keyphrases. Sun et al., SIGIR 2019

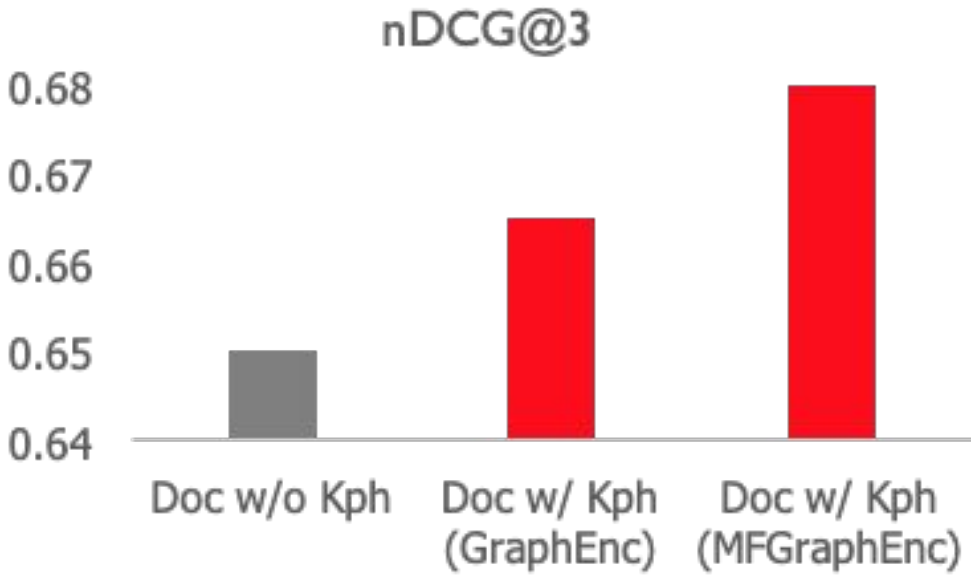
Experiment - Evaluation

- We evaluate models on **document ranking** task.



Experiment - Evaluation Result

- RQ 1. Whether **extracted keyphrases** improve document ranking.



Experiment - Evaluation Result

- RQ 2. Whether **leveraging structures** further improves performance.