

Web Document Encoding for Structure-Aware Keyphrase Extraction

Jihyuk Kim

Yonsei Univ



YONSEI
UNIVERSITY

Young-In Song

Naver Corp

NAVER

Seung-won Hwang

Seoul National Univ



SEOUL
NATIONAL
UNIVERSITY

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**.

NAVER

search



yoast.com > what-is-keyphrase-density-and-why-is-it-i
What is keyphrase density and why is it important?

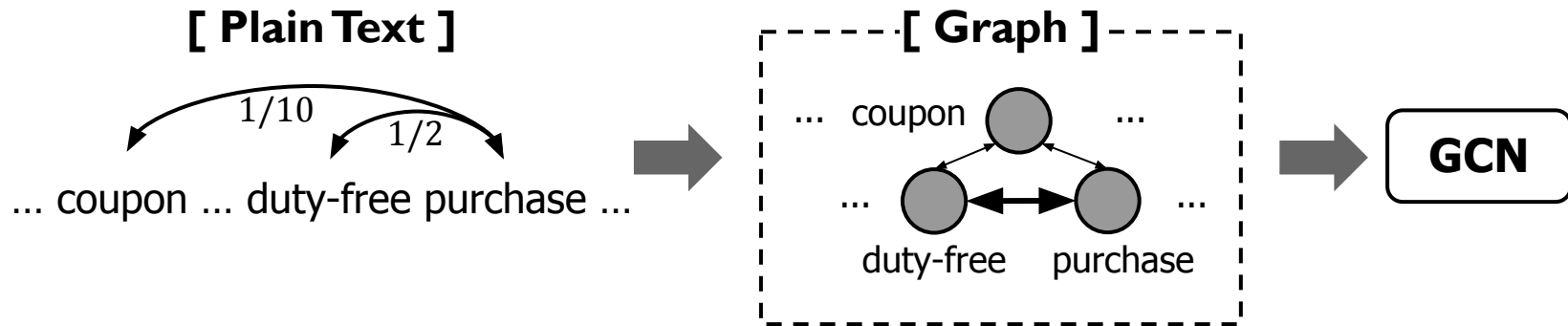
Why is your focus keyphrase density important? What does it
chec

www.yourdictionary.com > keyphrase
Keyphrase Meaning | Best 2 Definitions of Keyphrase

What does **keyphrase** mean? (cryptography) A phrase used in enc

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.
 - **Graph Convolutional Network (GCN)** is adopted to contextualize the graph.



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

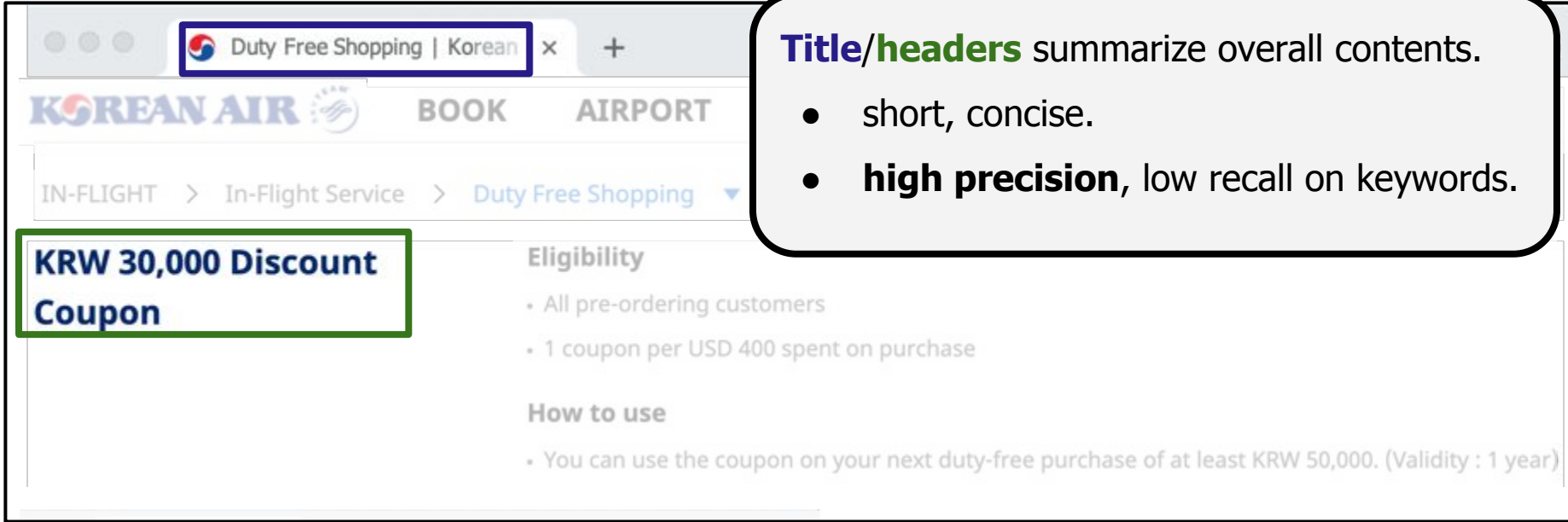
Motivation: Structure in Web Document, beyond Plain Text

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

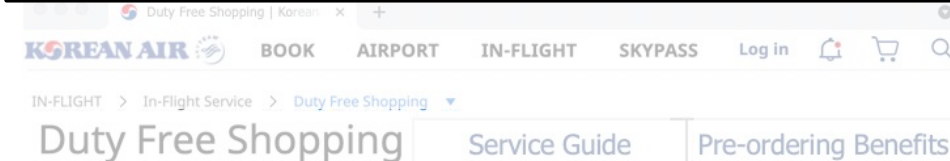
The image shows a screenshot of the Korean Air website's 'Duty Free Shopping' page. The browser tab is titled 'Duty Free Shopping | Korean'. The page header includes the Korean Air logo and navigation links for 'BOOK', 'AIRPORT', 'IN-FLIGHT', and 'SKYPASS', along with 'Log in', a notification bell, a shopping cart, and a search icon. The breadcrumb trail reads 'IN-FLIGHT > In-Flight Service > Duty Free Shopping'. The main content area is divided into two sections: a coupon and its details. The coupon, titled 'KRW 30,000 Discount Coupon', is highlighted with a green border. The details section, highlighted with a purple border, includes 'Eligibility' (All pre-ordering customers, 1 coupon per USD 400 spent) and 'How to use' (You can use the coupon on your next duty-free purchase of at least KRW 50,000). At the bottom, a navigation bar shows 'Duty Free Shopping', 'Service Guide', and 'Pre-ordering Benefits', with the latter highlighted by a red box and a red arrow pointing to it from the left. A red arrow also points to the browser tab. A cursor icon is visible at the bottom right.

Motivation: Multiple Fields with Complementary Benefits

- Those fields provide **complementary benefits**.



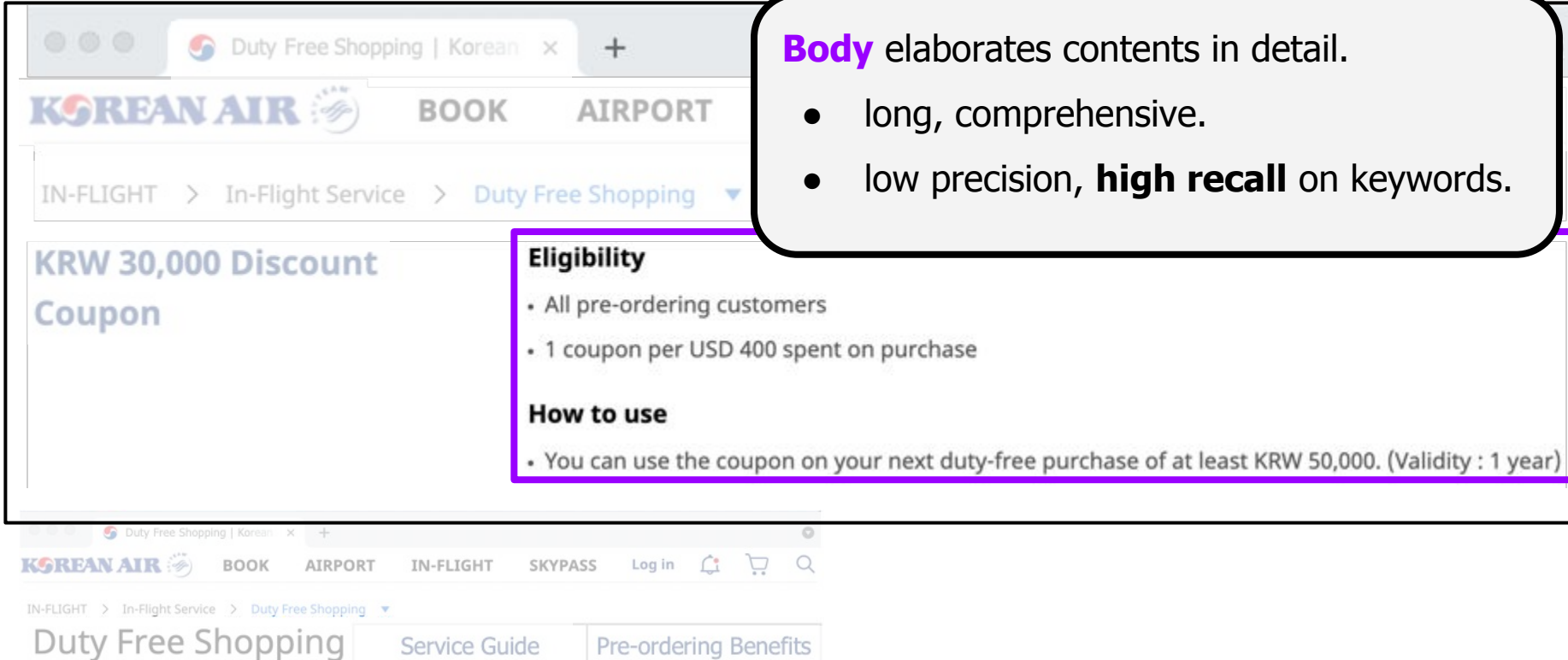
The screenshot shows a web browser window with the URL 'Duty Free Shopping | Korean'. The page header includes the Korean Air logo and navigation links for 'BOOK' and 'AIRPORT'. The breadcrumb trail is 'IN-FLIGHT > In-Flight Service > Duty Free Shopping'. A green box highlights the main heading: 'KRW 30,000 Discount Coupon'. To the right, under the heading 'Eligibility', there are two bullet points: 'All pre-ordering customers' and '1 coupon per USD 400 spent on purchase'. Under the heading 'How to use', there is one bullet point: 'You can use the coupon on your next duty-free purchase of at least KRW 50,000. (Validity : 1 year)'. A callout box on the right contains the text: 'Title/headers summarize overall contents.' followed by two bullet points: 'short, concise.' and 'high precision, low recall on keywords.'



The screenshot shows the bottom part of the Korean Air website. The breadcrumb trail is 'IN-FLIGHT > In-Flight Service > Duty Free Shopping'. Below the breadcrumb, there are two buttons: 'Duty Free Shopping' and 'Service Guide'. At the bottom, there are navigation links for 'BOOK', 'AIRPORT', 'IN-FLIGHT', 'SKYPASS', 'Log in', a bell icon, a shopping cart icon, and a search icon.

Motivation: Multiple Fields with Complementary Benefits

- Those fields provide **complementary benefits**.



Body elaborates contents in detail.

- long, comprehensive.
- low precision, **high recall** on keywords.

KRW 30,000 Discount Coupon

Eligibility

- All pre-ordering customers
- 1 coupon per USD 400 spent on purchase

How to use

- You can use the coupon on your next duty-free purchase of at least KRW 50,000. (Validity : 1 year)

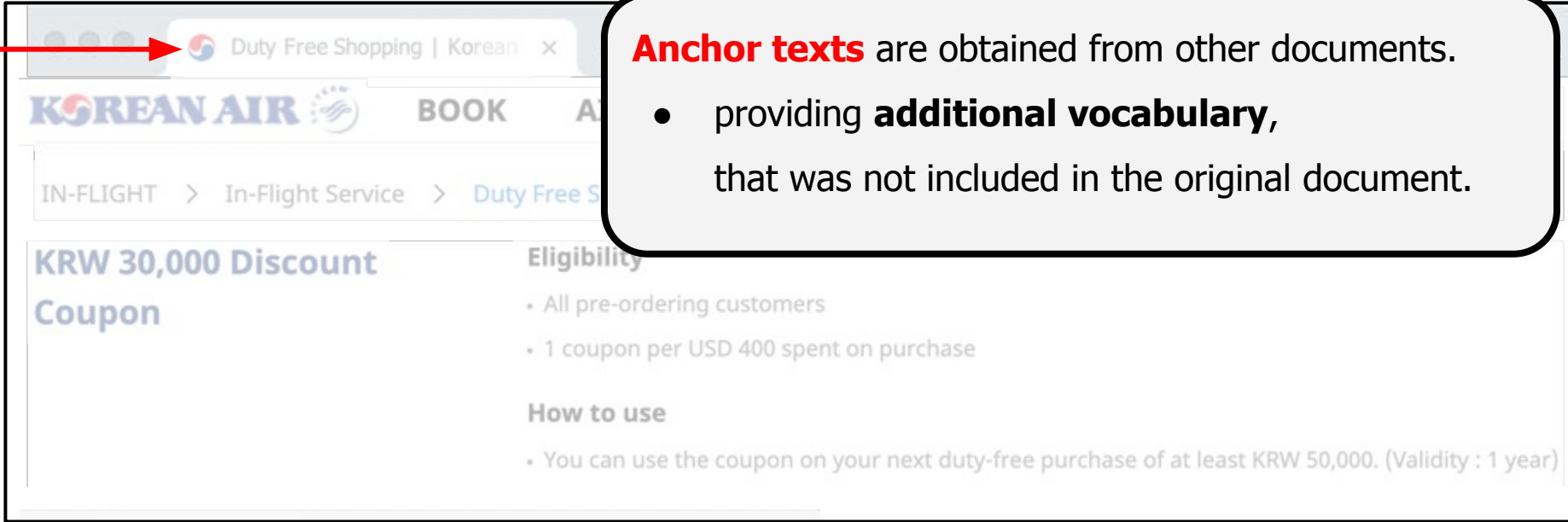
KOREAN AIR BOOK AIRPORT IN-FLIGHT SKYPASS Log in

IN-FLIGHT > In-Flight Service > Duty Free Shopping

Duty Free Shopping Service Guide Pre-ordering 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.

The screenshot shows the Korean Air website with a coupon for KRW 30,000. The coupon details include eligibility (All pre-ordering customers, 1 coupon per USD 400 spent) and how to use it (on a next duty-free purchase of at least KRW 50,000). A red arrow points from the browser tab to a callout box containing the text above.

Duty Free Shopping Service Guide **Pre-ordering Benefits**



Motivation: Multiple Fields with Complementary Benefits

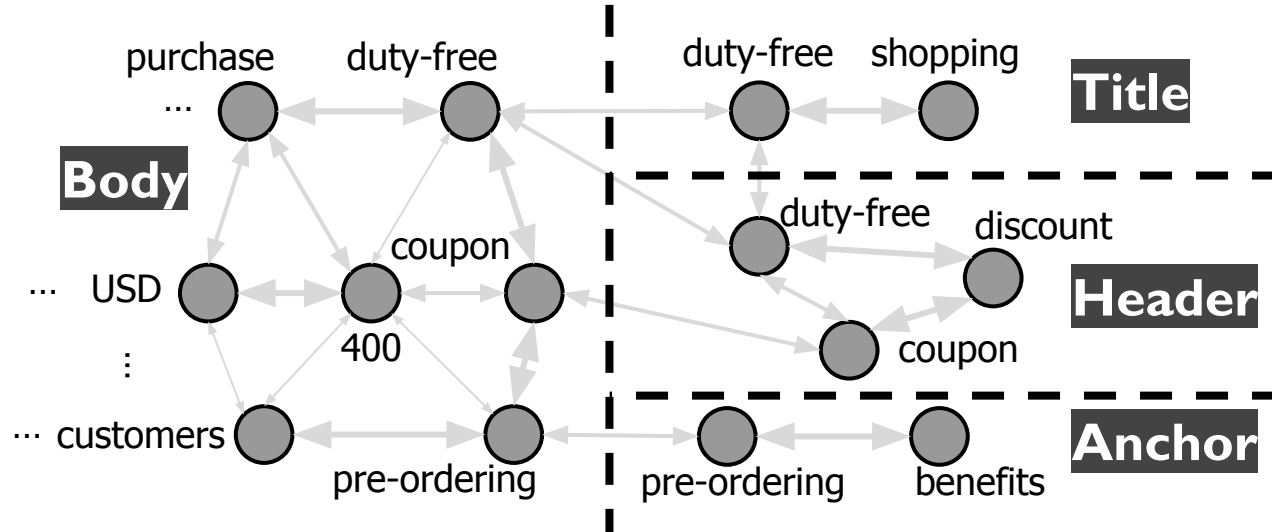
- Those fields provide **complementary benefits**.

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.

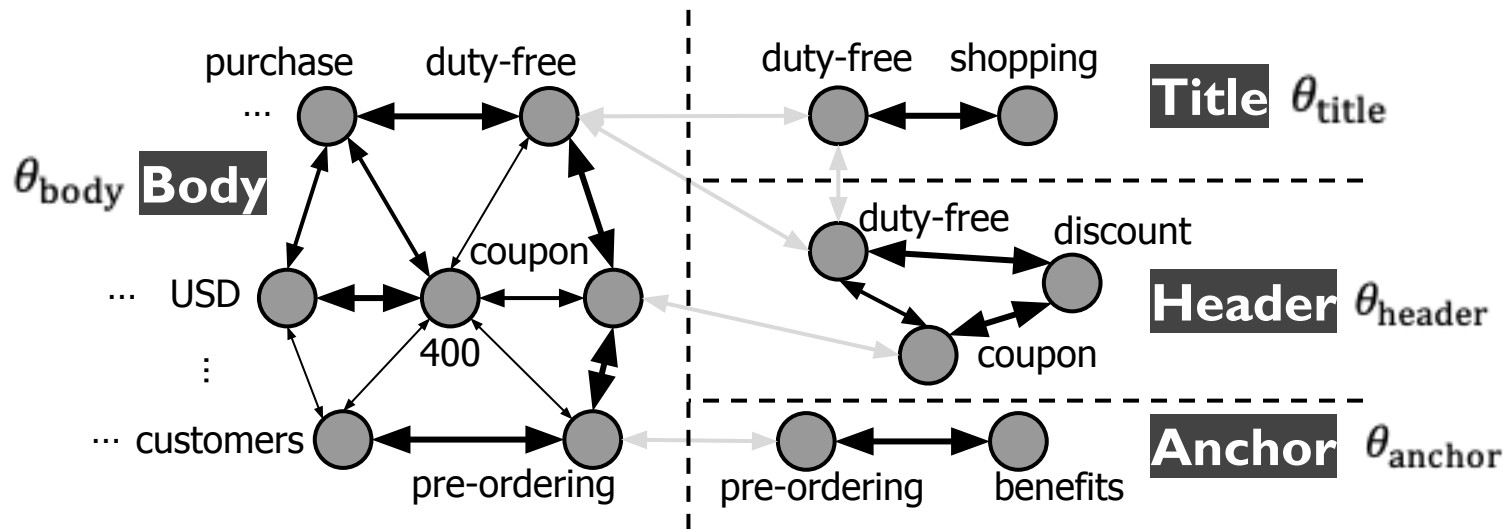
Approach: Multi-Field Graph Encoding

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



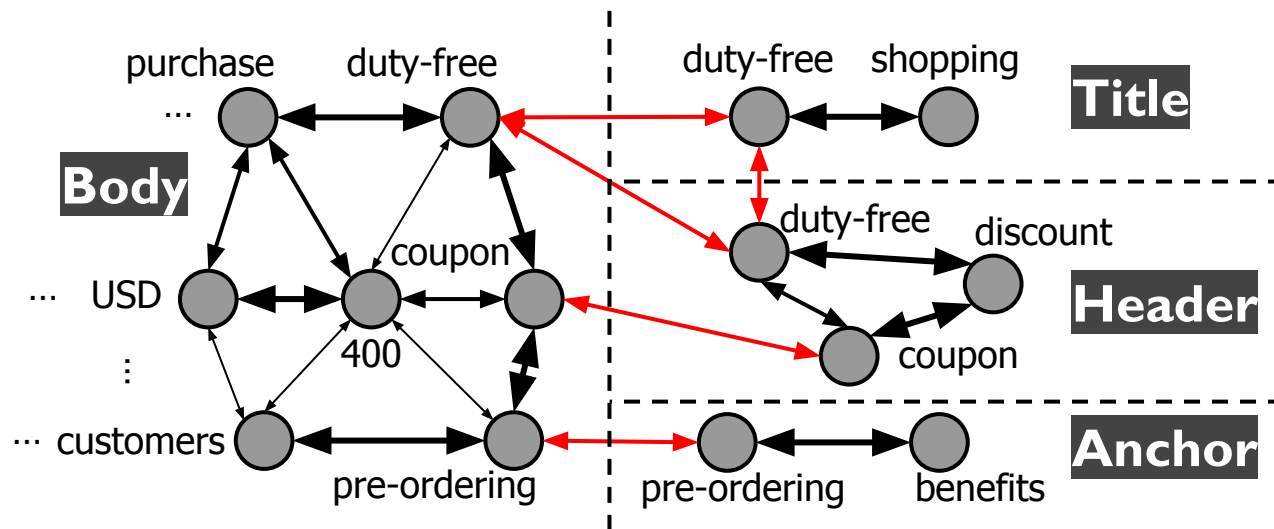
Approach: Multi-Field Graph Encoding

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



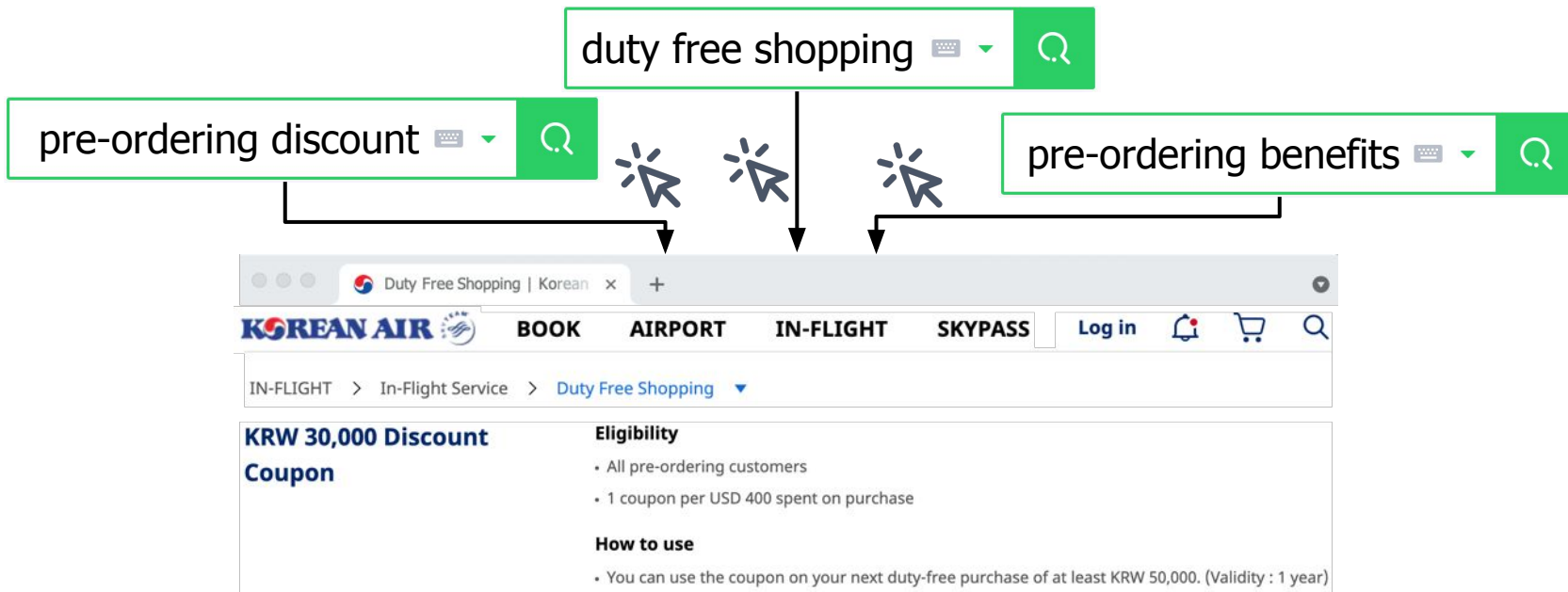
Approach: Multi-Field Graph Encoding

- We represent multiple field contents using an unified graph.
 - **Inter-field** edges (\longleftrightarrow) 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

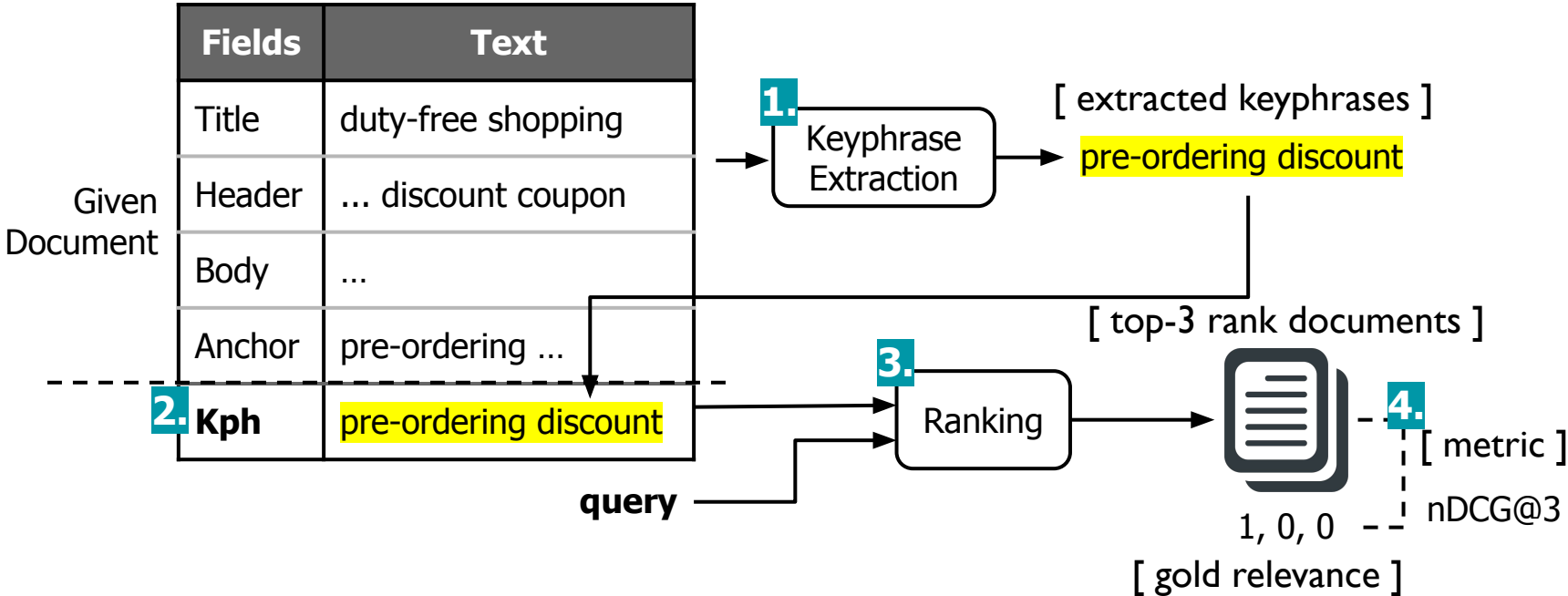
- We compare encoders with and without structures.

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

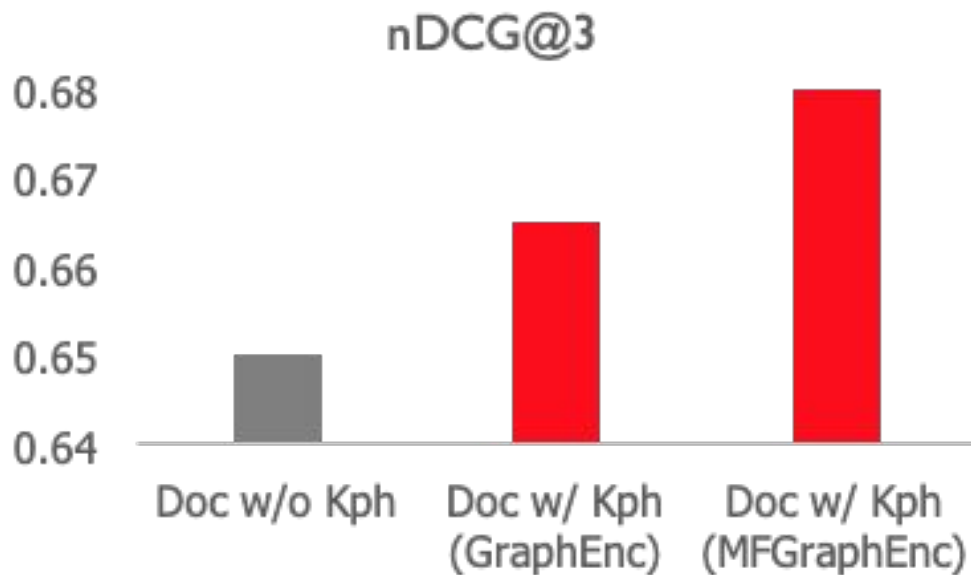
Experiment - Evaluation

- 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.

