Heading-aware Snippet Generation for Web Search

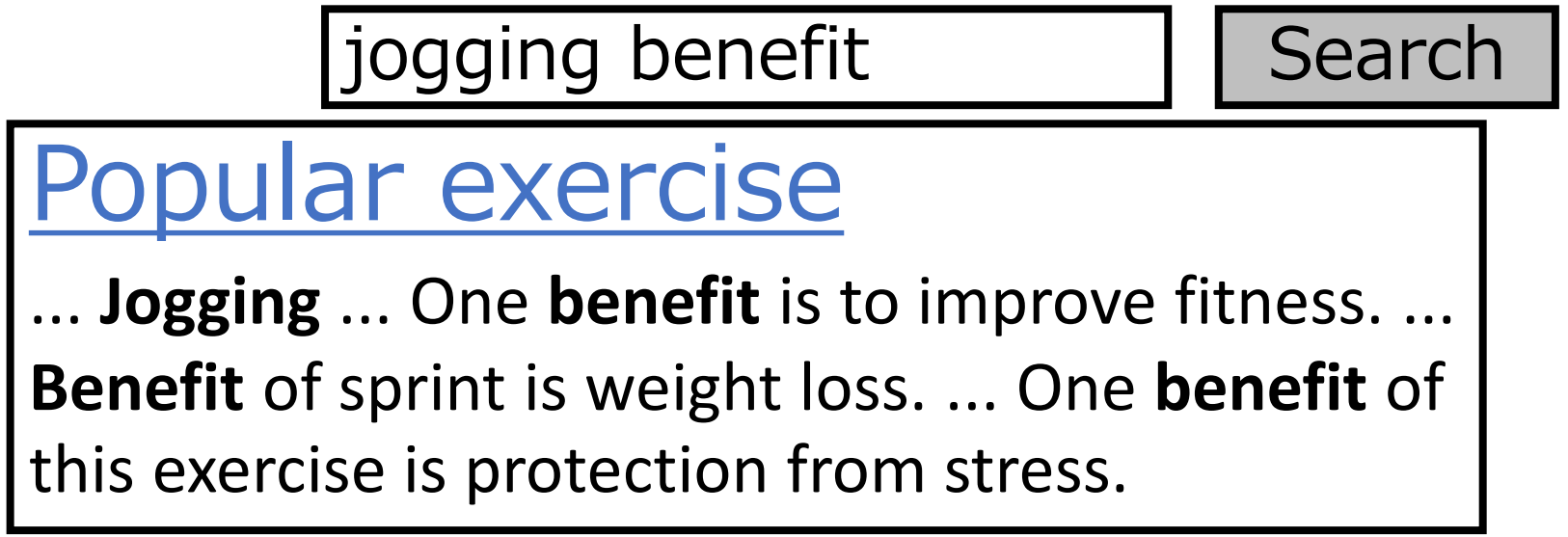
Tomohiro Manabe and Keishi Tajima

Graduate School of Informatics, Kyoto Univ.

{manabe@dl.kuis, tajima@i}.kyoto-u.ac.jp

Web Search Result Snippets

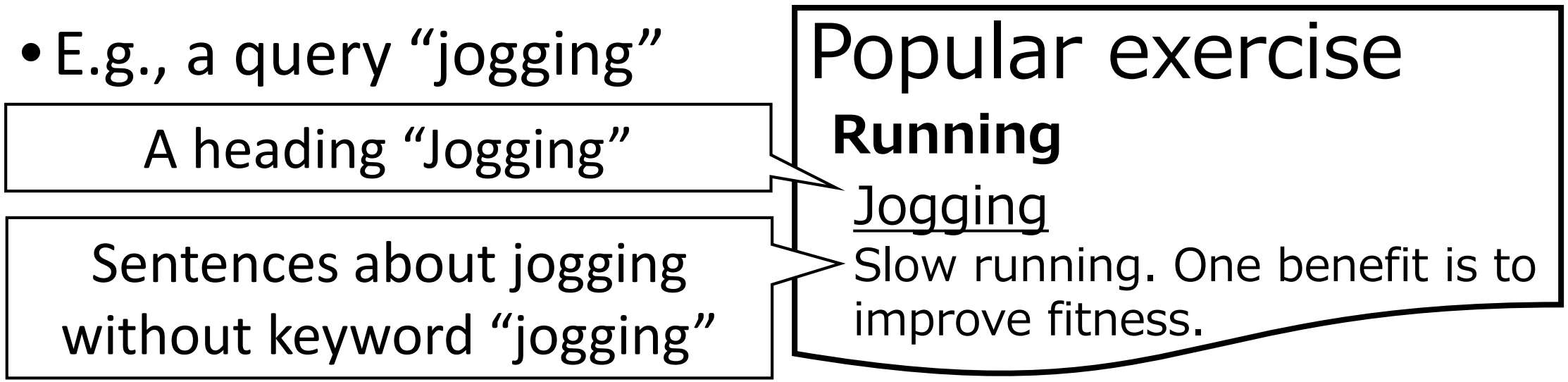
- Are short summaries of web page text



- Search engine users read them and judge relevance of original pages to search intents

Key Idea

- To generate snippets, search engines rank sentences
- *Headings* of sentences are important to rank them
- E.g., a query “jogging”



- We propose heading-aware generation methods

Key Idea

- To generate snippets, search engines rank sentences
- *Headings* of sentences are important to rank them
- E.g., a query “jogging”

A heading “Jogging”

Sentences about jogging without keyword “jogging”

Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

- We propose heading-aware generation methods

Key Idea

- To generate snippets, search engines rank sentences
- *Headings* of sentences are important to rank them
- E.g., a query “jogging”

A heading “Jogging”

Sentences about jogging without keyword “jogging”

Popular exercise

Running

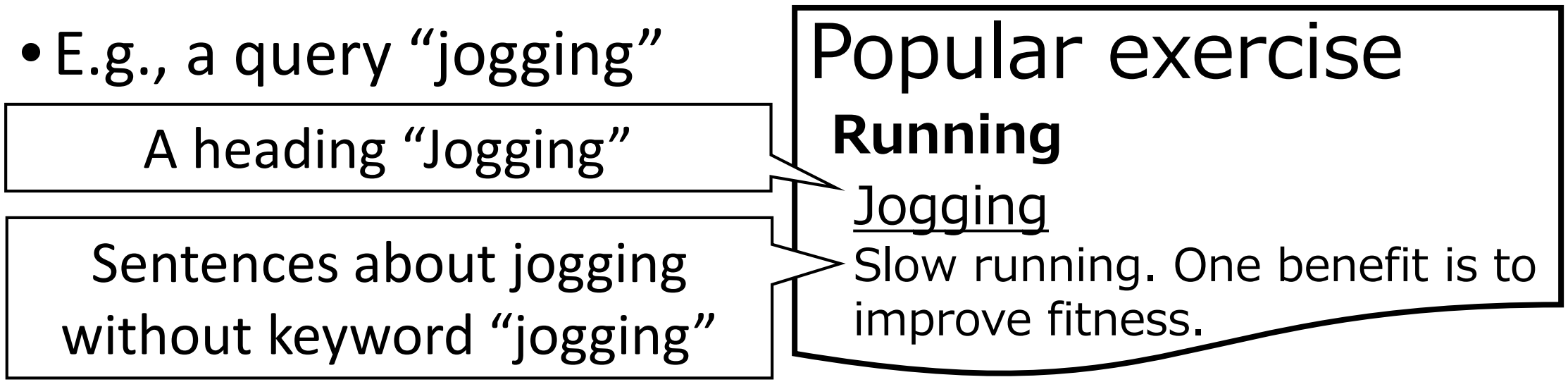
Jogging

Slow running. One benefit is to improve fitness.

- We propose heading-aware generation methods

Key Idea

- To generate snippets, search engines rank sentences
- *Headings* of sentences are important to rank them
- E.g., a query “jogging”



- We propose heading-aware generation methods

Outline of This Presentation

I. Our definition of hierarchical heading structure
II. Heading-aware snippet generation methods
III. Current evaluation result

Our definition of hierarchical heading structure

Hierarchical heading structure and its components

In Short

- *Hierarchical Heading Structure* is composed of
  - nested logical *blocks*
  - associated with *headings*
- Each heading describes a block topic briefly

Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

Sprint

Benefit of sprint is weight loss.

Swimming

Front Crawl

One benefit of this exercise is protection from stress.

In Short

- *Hierarchical Heading Structure* is composed of
  - nested logical *blocks*
  - associated with *headings*
- Each heading describes a block topic briefly

Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

Sprint

Benefit of sprint is weight loss.

Swimming

Front Crawl

One benefit of this exercise is protection from stress.

A *Heading* is

- A highly summarized topic description for a part of a web page
- *Heading words* are words in headings

Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

Sprint

Benefit of sprint is weight loss.

Swimming

Front Crawl

One benefit of this exercise is protection from stress.

A *Heading* is

- A highly summarized topic description for a part of a web page
- *Heading words* are words in headings

Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

Sprint

Benefit of sprint is weight loss.

Swimming

Front Crawl

One benefit of this exercise is protection from stress.

A *block* is

- A part of a web page
  - associated with a heading
- Note that
  - An entire web page is also a block
  - There is one-to-one correspondence between headings and blocks

Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

Sprint

Benefit of sprint is weight loss.

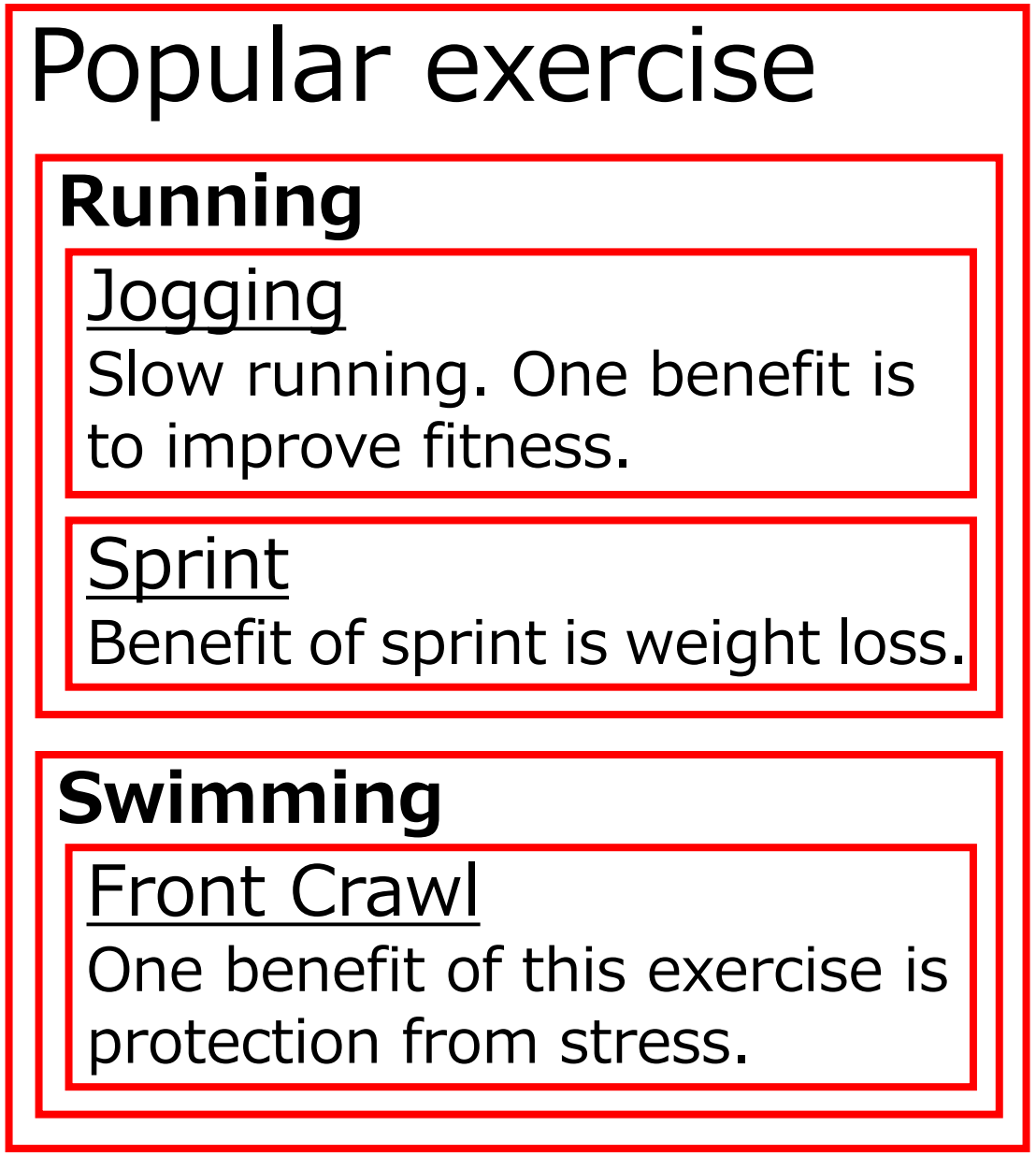
Swimming

Front Crawl

One benefit of this exercise is protection from stress.

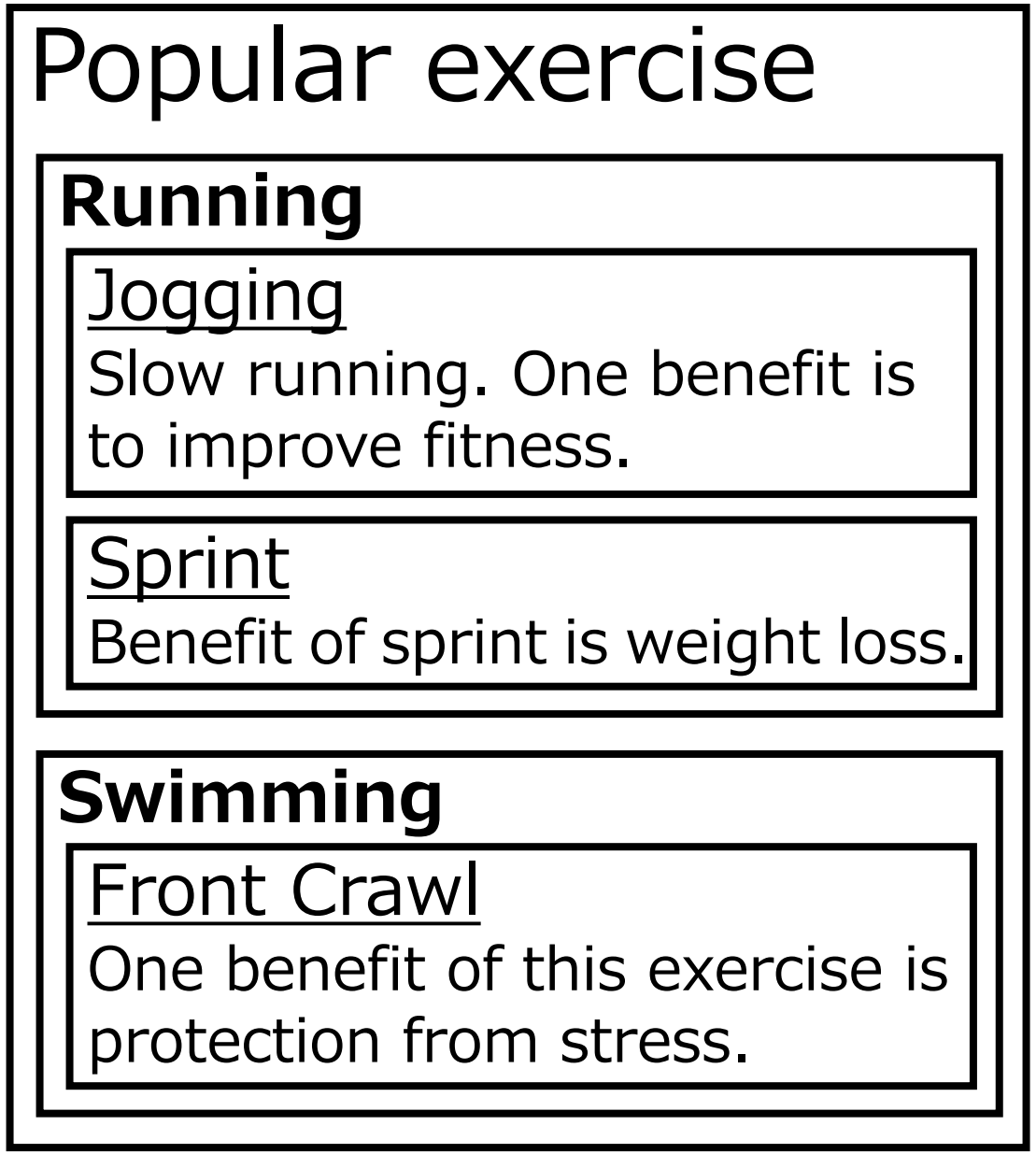
A *block* is

- A part of a web page
  - associated with a heading
- Note that
  - An entire web page is also a block
  - There is one-to-one correspondence between headings and blocks



Hierarchical Heading Structure

- A block may include other blocks entirely
- Blocks in a page form hierarchical heading structure
  - Its root is the entire page
- Our methods focus on such structure



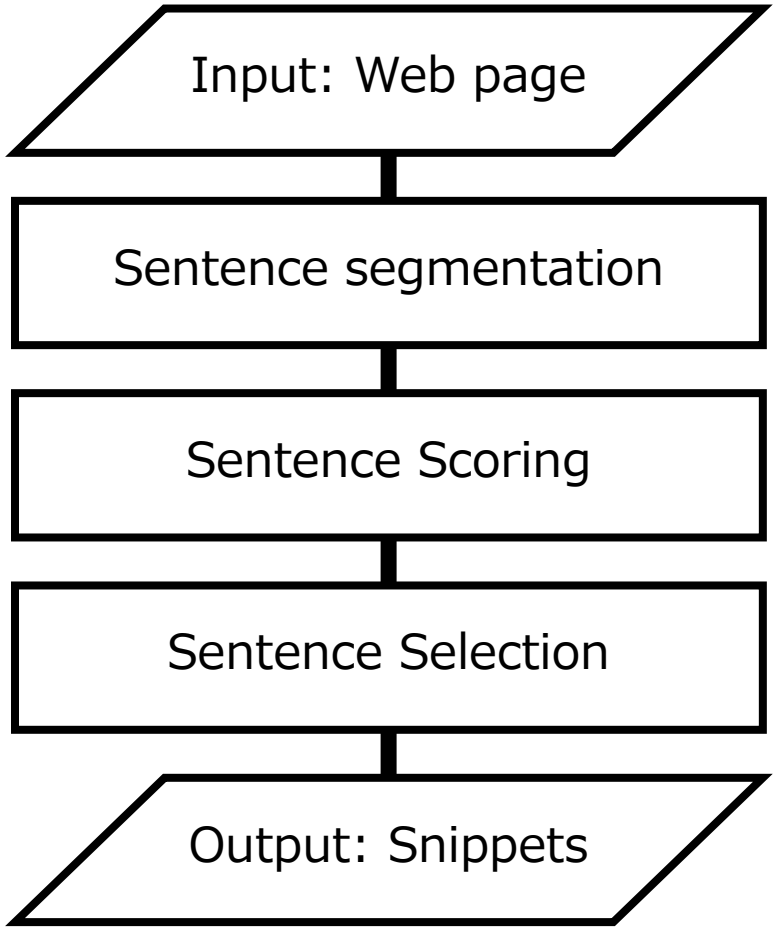
Hierarchical Heading Structure Extraction from Web Pages

- is NOT a trivial problem
- Throughout this presentation, the structure is given
- For evaluation, we used previously proposed method
  - Its implementation is at https://github.com/tmanabe

Manabe, T., Tajima, K.: Extracting logical hierarchical structure of HTML documents based on headings. VLDB 8(12), 1606–1617 (2015)

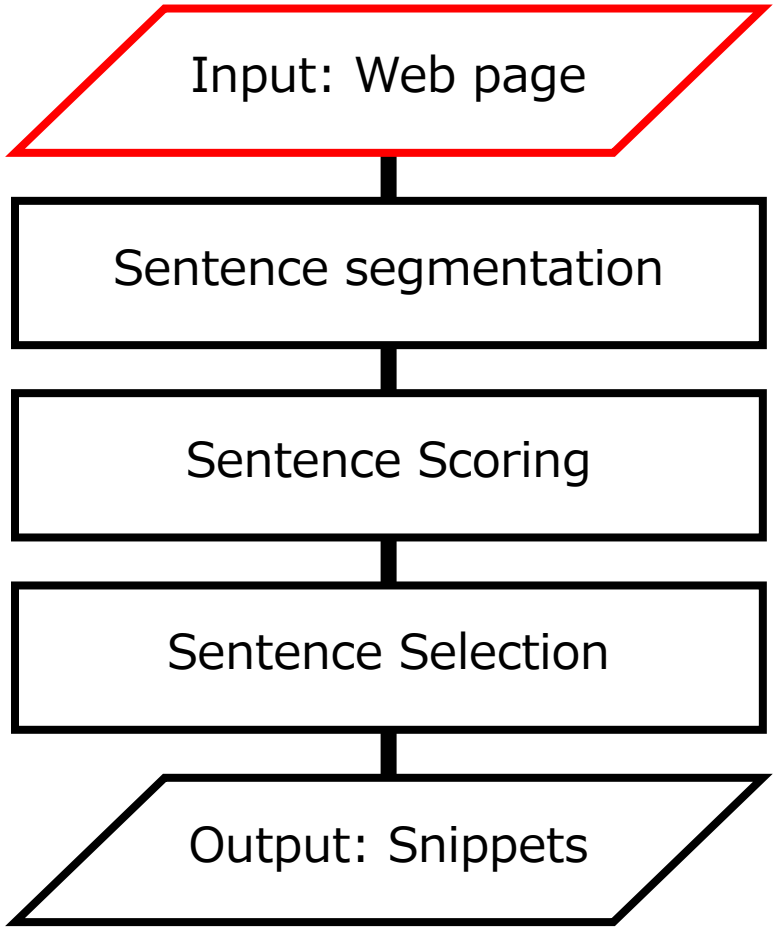
Heading-aware Snippet Generation Methods

Basic Snippet Generation Method



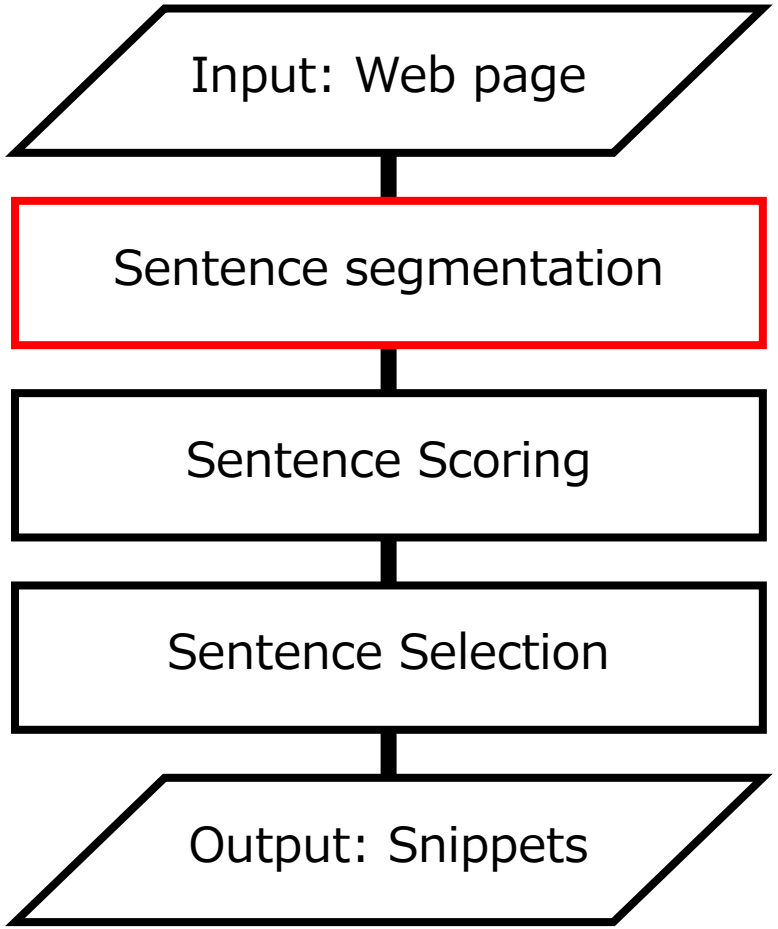
1. Split the page into semantically coherent fragments (e.g. sentences)
2. Score the fragments
   - By using scoring functions based on query keyword occurrences
   - E.g., TFIDF and BM25
3. Select the fragments
   - In desc. order of their scores
   - Until the snippet length reaches limit

Basic Snippet Generation Method



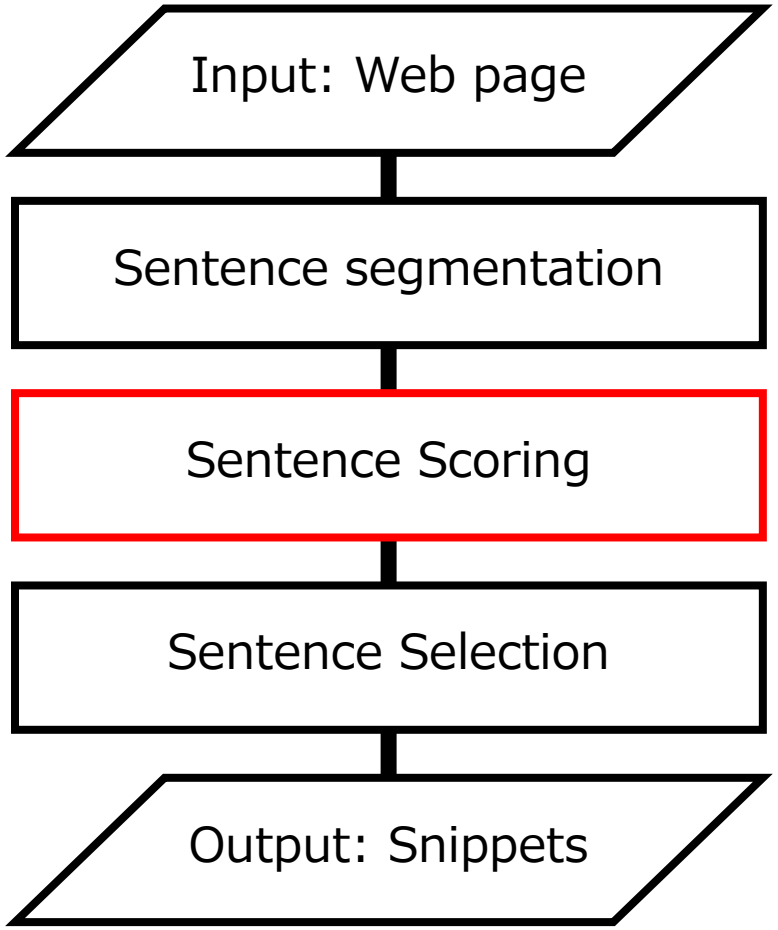
1. Split the page into semantically coherent fragments (e.g. sentences)
2. Score the fragments
   - By using scoring functions based on query keyword occurrences
   - E.g., TFIDF and BM25
3. Select the fragments
   - In desc. order of their scores
   - Until the snippet length reaches limit

Basic Snippet Generation Method



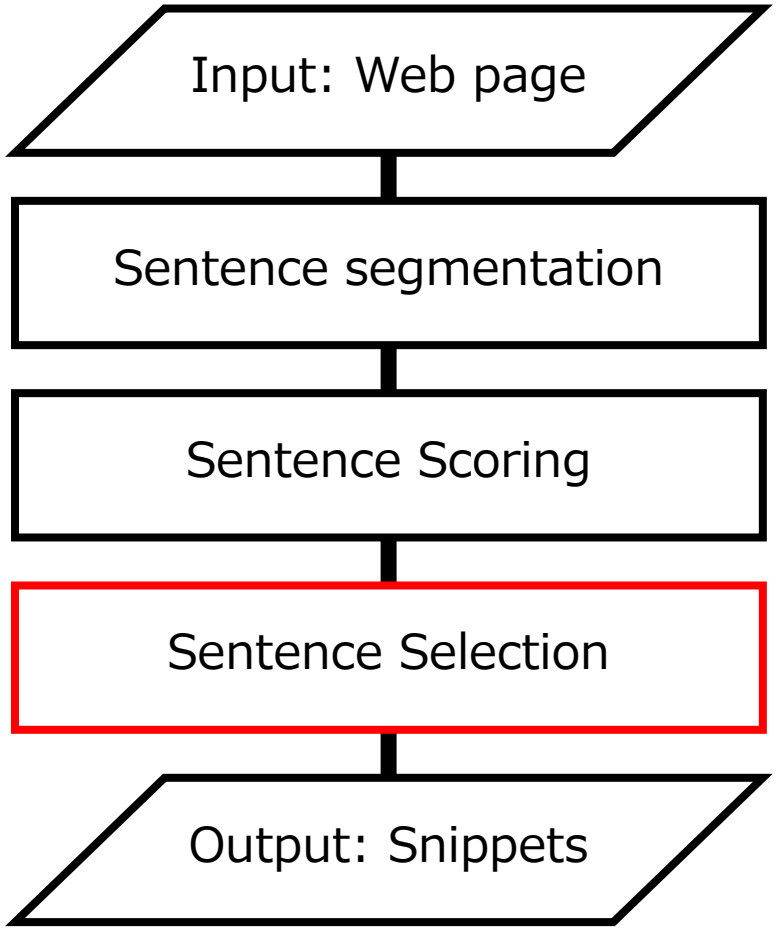
1. Split the page into semantically coherent fragments (e.g. sentences)
2. Score the fragments
   - By using scoring functions based on query keyword occurrences
   - E.g., TFIDF and BM25
3. Select the fragments
   - In desc. order of their scores
   - Until the snippet length reaches limit

Basic Snippet Generation Method



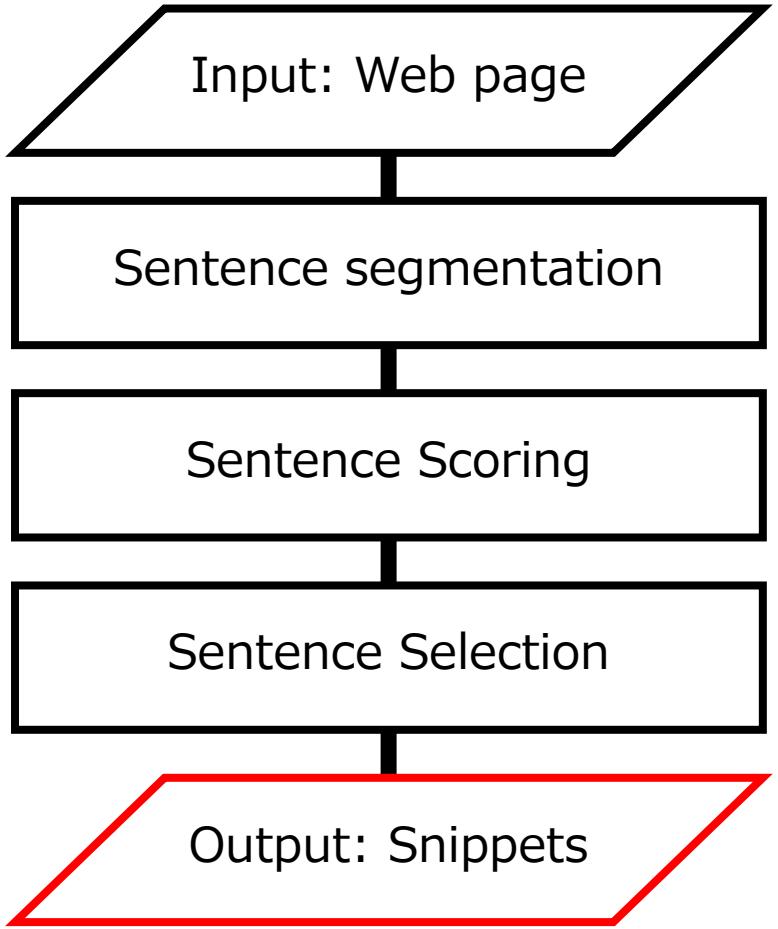
1. Split the page into semantically coherent fragments (e.g. sentences)
2. Score the fragments
   - By using scoring functions based on query keyword occurrences
   - E.g., TFIDF and BM25
3. Select the fragments
   - In desc. order of their scores
   - Until the snippet length reaches limit

Basic Snippet Generation Method



1. Split the page into semantically coherent fragments (e.g. sentences)
2. Score the fragments
   - By using scoring functions based on query keyword occurrences
   - E.g., TFIDF and BM25
3. Select the fragments
   - In desc. order of their scores
   - Until the snippet length reaches limit

Basic Snippet Generation Method



1. Split the page into semantically coherent fragments (e.g. sentences)
2. Score the fragments
   - By using scoring functions based on query keyword occurrences
   - E.g., TFIDF and BM25
3. Select the fragments
   - In desc. order of their scores
   - Until the snippet length reaches limit

Four Methods

- All follow these steps

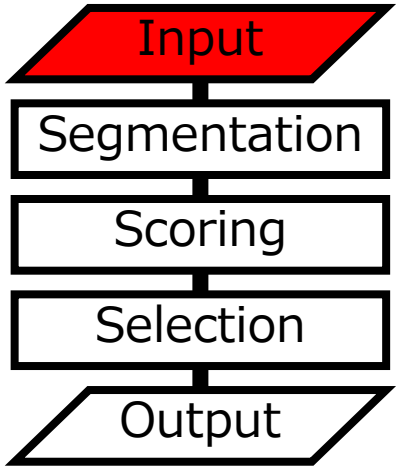
1. Baseline method
   - Query keywords in sentence indicate importance
2. *Existing* method
   - Heading words in sentence also indicate importance
3. *Our* method
   - Query keywords in headings also indicate importance
4. *Combined* method
   - All three ideas

Four Methods

- All follow these steps

1. Baseline method
   - Query keywords in sentence indicate importance
2. *Existing* method
   - Heading words in sentence also indicate importance
3. *Our* method
   - Query keywords in headings also indicate importance
4. *Combined* method
   - All three ideas

1. Baseline Method



- Input
  - A web page

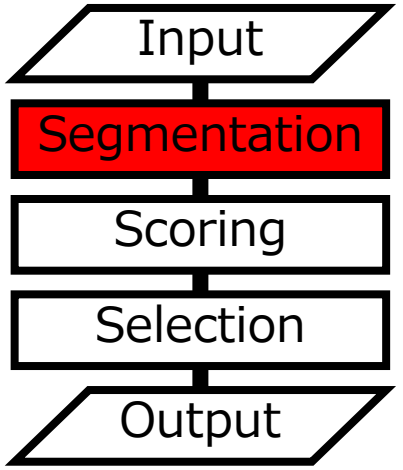
Popular exercise

Running

Jogging

Slow running. One benefit is to improve fitness.

1. Baseline Method



- First step
  - The method segments the page into sentences
  - NOT the main topic of our research

Popular exercise

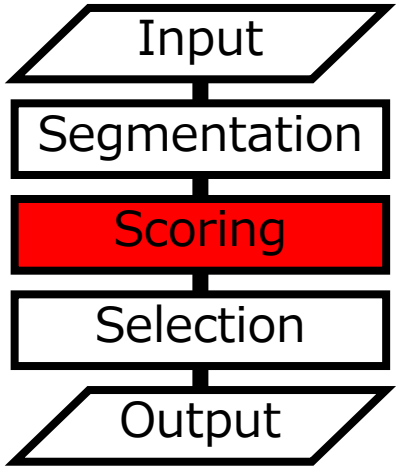
Running

Jogging

Slow running. One benefit is to improve fitness.

Popular exercise
Running
Jogging
Slow running.
One benefit is to improve fitness.
…

1. Baseline Method



- Second step
  - The method scores the sentences based on the number of query keywords in them
  - The main topic of our research
  - But as the baseline, we used

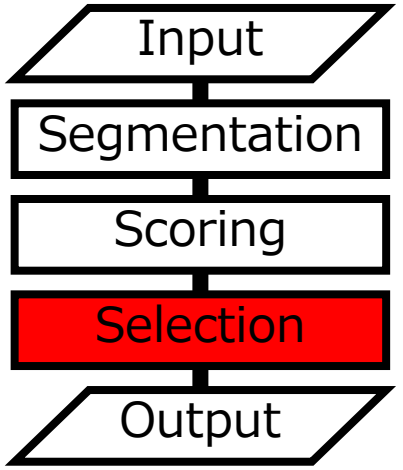
$$\mathrm{BM25}(\text{Query keywords})$$

jogging benefit | Search

Popular exercise
Running
Jogging
Slow running.
One benefit is to improve fitness.
…

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: TREC. pp. 109-126 (1996)

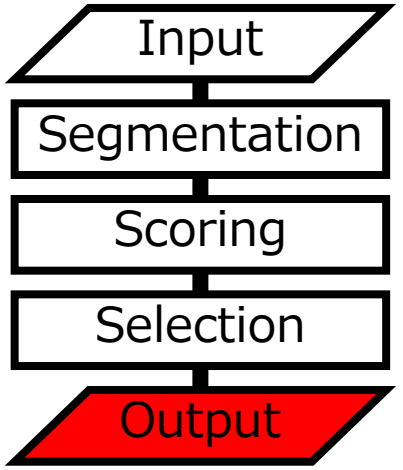
1. Baseline Method



- Third step
  - The method selects the sentences
  - It simply scans the sentences in desc. order of score
  - If there remains space to include sentence, it does so
  - NOT the main topic of our research

Jogging
One benefit is to improve fitness.
Benefit of sprint is weight loss.
One benefit of this exercise is protection from stress.
…

1. Baseline Method



- Output
  - Title (or URL if there is no title)
  - Snippets

Popular exercise

... **Jogging** ... One **benefit** is to improve fitness. ... **Benefit** of sprint is weight loss. ... One **benefit** of this exercise is protection from stress.

Four Methods

1. Baseline method
   - Query keywords in sentence indicate importance
2. *Existing* method
   - Heading words in sentence also indicate importance
3. *Our* method
   - Query keywords in headings also indicate importance of sentences in their associated blocks
4. *Combined* method
   - All three ideas

2. Existing Method

- An existing idea of Pembe and Güngör
  - Heading words in sentences indicate importance
- Because the heading words are expected to be important words in the block

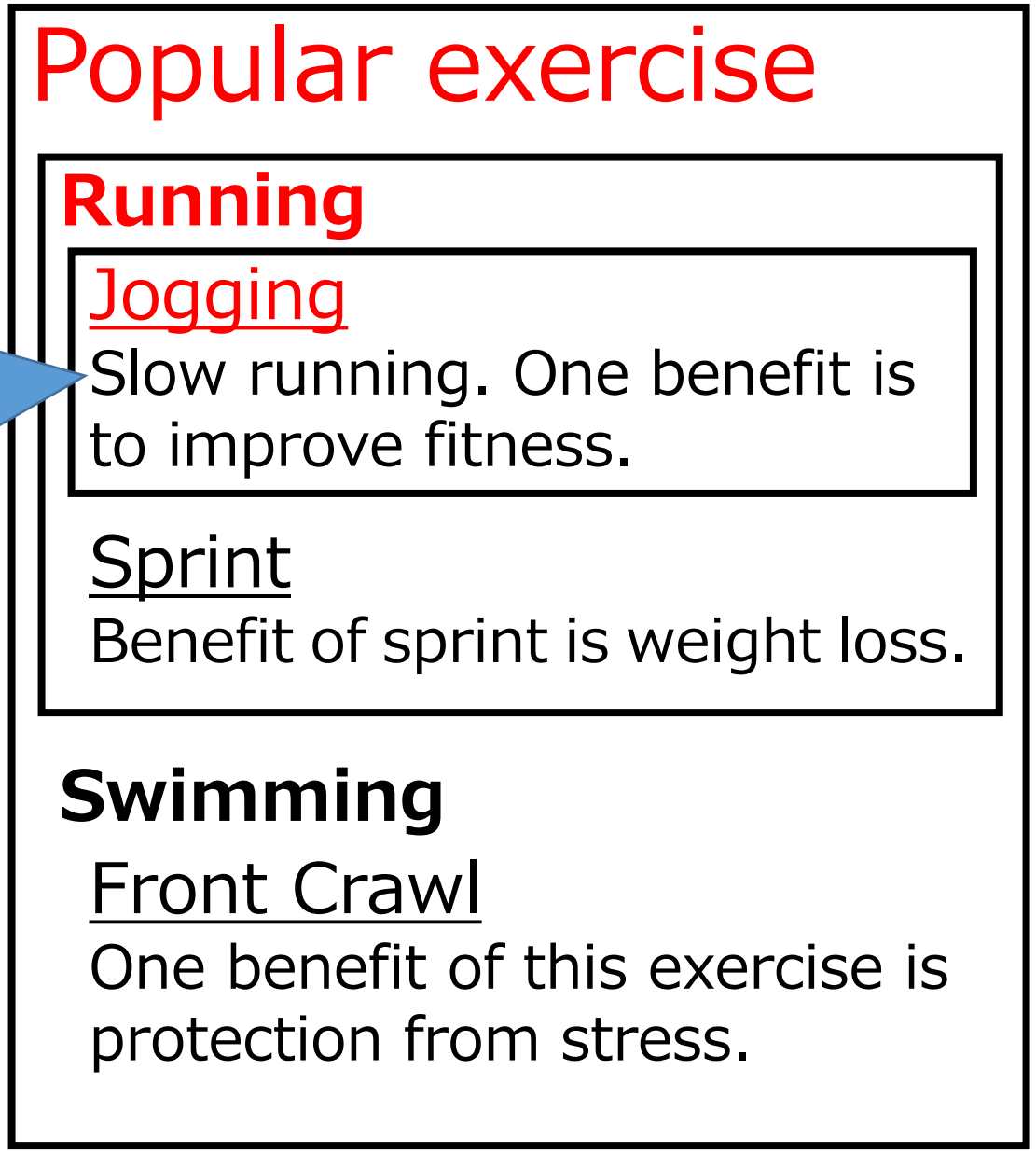
Pembe, F.C., Güngör, T.: Structure-preserving and query-biased document summarization for web searching. Online Info. Rev. 33(4), 696-719 (2009)

2. Existing Method

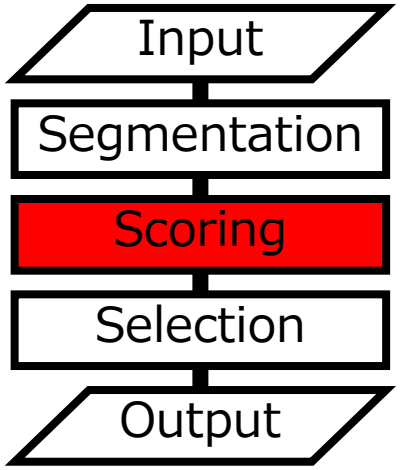
- E*xisting* method
  - counts heading words in sentences to score the sentences
  - because heading words are important for the sentences
- Heading words of a sentence
  - are words in hierarchical headings of the block including it
  - Hierarchical headings: headings of ancestor-or-self blocks

Heading Words of a Sentence

- Heading words of "Slow running." are:
  - Popular, exercise, running, and jogging
- The sentence includes a heading word "running"
- It is important



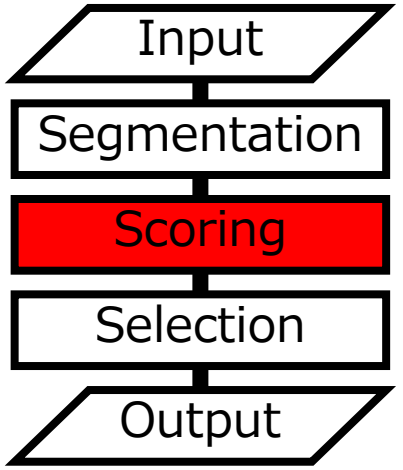
2. Existing Method



- The scoring function for sentences
  - Simply,
    $\mathrm{BM25}(\text{Query keywords}) + \mathrm{BM25}(\text{Heading words})$
  - Problem
    - Summation of BM25 scores produces worse ranking when they count the same words
    - In other words,
      when heading words include some query keywords

Robertson, S., Zaragoza, H., Taylor, M.:
Simple BM25 extension to multiple weighted fields. In: CIKM. pp. 42-49 (2004)

2. Existing Method



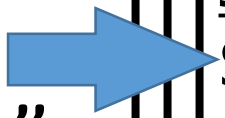
- Therefore, we split words into three types

	Not query keywords	Query keywords
Not heading words	(Do not care)	Query-only words
Heading words	Heading-only words	Query-heading words

- Sum up their scores
- I.e., $\mathrm{BM25}(\text{Query only}) + \mathrm{BM25}(\text{Heading only}) + \mathrm{BM25}(\text{Query heading})$

Three Word Types

- For “Slow running.” and a query “jogging benefit”
  - Query-heading word is jogging
  - Heading-only words are popular, exercise, running
  - Query-only word is benefit



Popular exercise

Running

Jogging
Slow running. One benefit is to improve fitness.

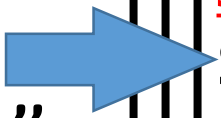
Sprint
Benefit of sprint is weight loss.

Swimming

Front Crawl
One benefit of this exercise is protection from stress.

Three Word Types

- For “Slow running.” and a query “jogging benefit”
  - Query-heading word is jogging
  - Heading-only words are popular, exercise, running
  - Query-only word is benefit



Popular exercise

Running

Jogging
Slow running. One benefit is to improve fitness.

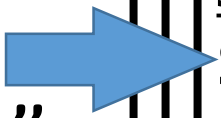
Sprint
Benefit of sprint is weight loss.

Swimming

Front Crawl
One benefit of this exercise is protection from stress.

Three Word Types

- For “Slow running.” and a query “jogging benefit”
  - Query-heading word is jogging
  - Heading-only words are popular, exercise, running
  - Query-only word is benefit



Popular exercise

Running

Jogging
Slow running. One benefit is to improve fitness.

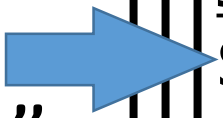
Sprint
Benefit of sprint is weight loss.

Swimming

Front Crawl
One benefit of this exercise is protection from stress.

Three Word Types

- For “Slow running.” and a query “jogging benefit”
  - Query-heading word is jogging
  - Heading-only words are popular, exercise, running
  - Query-only word is benefit



Popular exercise

Running

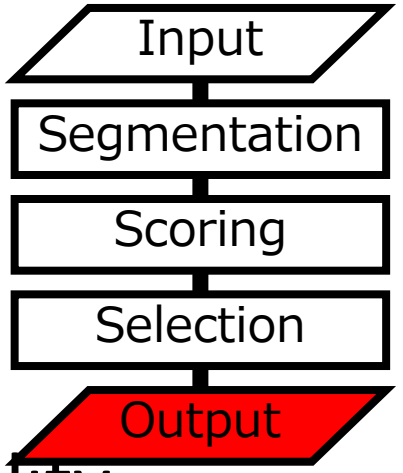
Jogging
Slow running. One benefit is to improve fitness.

Sprint
Benefit of sprint is weight loss.

Swimming

Front Crawl
One benefit of this exercise is protection from stress.

2. Existing Method



- Modified output
  - Showed headings separately to improve readability

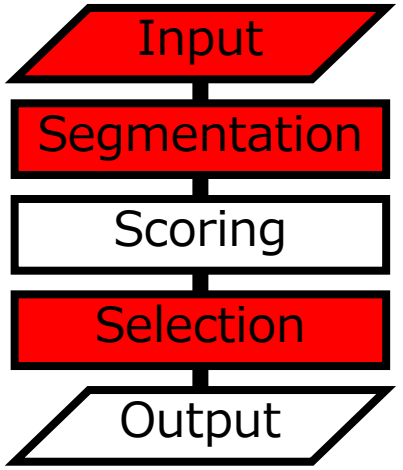
> Popular exercise
>
> **> Running > Sprint**
>
> Benefit of sprint is weight loss.
>
> **> Swimming > Front Crawl**
>
> One benefit of this exercise is protection from stress.

  - Headings shown iff sentences in their blocks are chosen

Pembe, F.C., Güngör, T.: Structure-preserving and query-biased document summarization for web searching. Online Info. Rev. 33(4), 696-719 (2009)

2. Existing Method

- Other steps are same as those of the baseline method

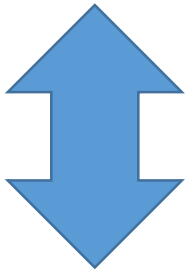


Four Methods

1. Baseline method
   - Query keywords in sentence indicate importance
2. *Existing* method
   - Heading words in sentence also indicate importance
3. *Our* method
   - Query keywords in headings also indicate importance of sentences in their associated blocks
4. *Combined* method
   - All three ideas

3. Our Method

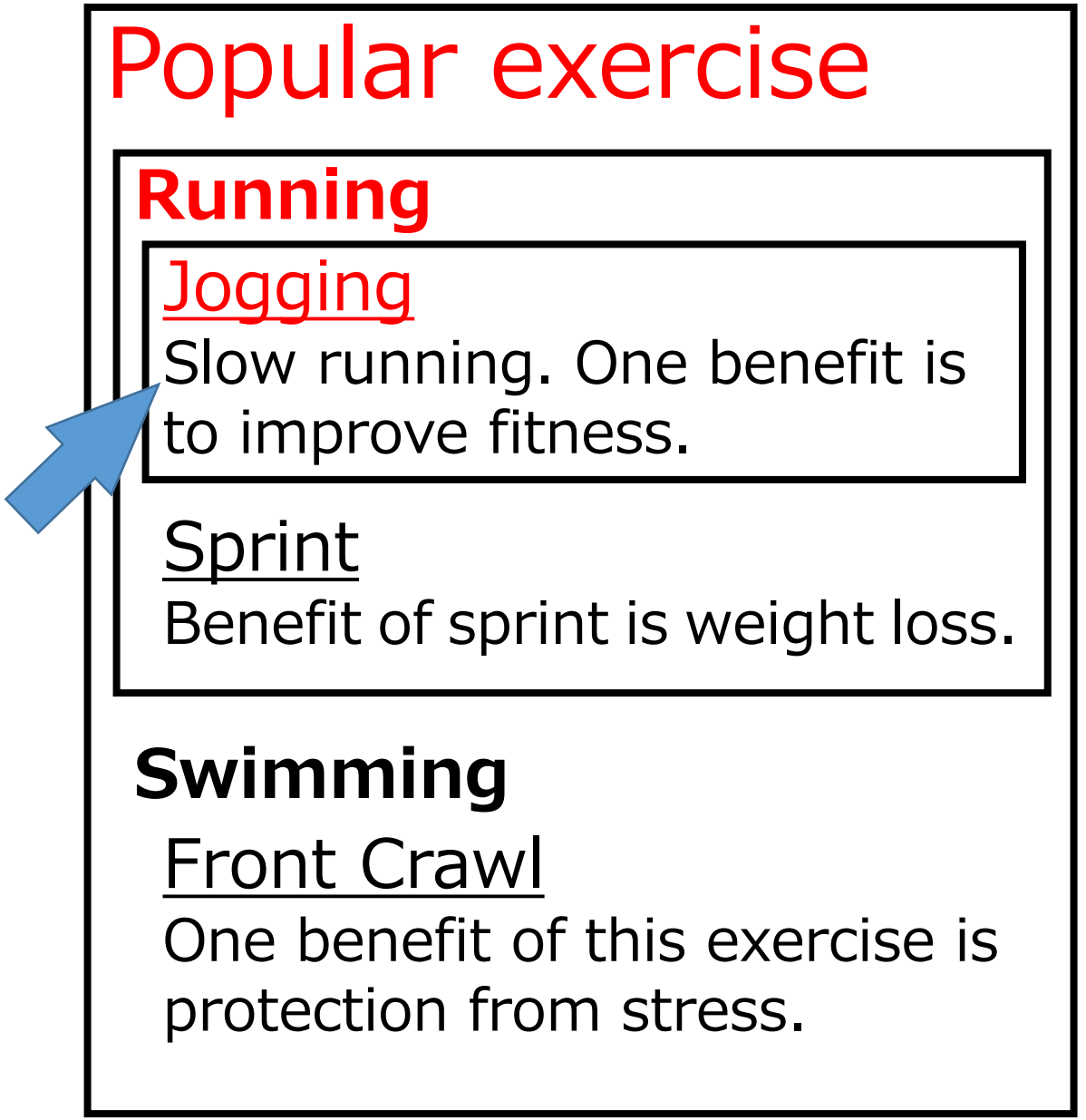
- The existing idea
  - Heading words in sentences indicate importance of them



- Our idea
  - Query keywords in a heading indicate importance of sentences in its associated blocks

Omission of Heading Words

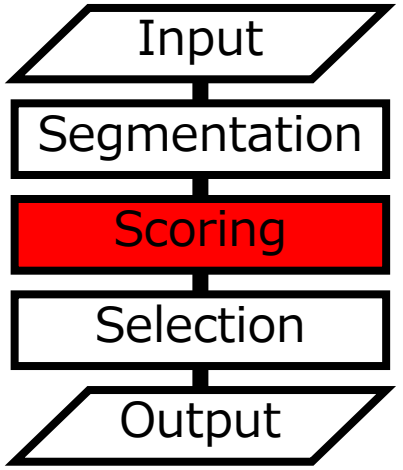
- Heading words are very often omitted
- Sentence “Slow running” is talking about popular exercise, running, and jogging but the heading words popular, exercise, and jogging are omitted



3. Our Method

- Takes the omission of heading words into account
- Assigns high scores to sentences including query keywords within either
  - Sentences themselves
  - Their hierarchical headings

3. Our Method

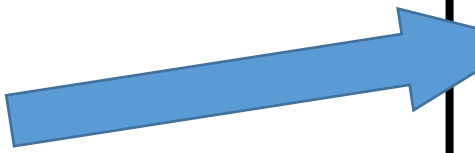


- Sentence scoring modified in different way
- Each sentence comprises <u>two fields</u>
  - Contents of the sentence itself
  - Its hierarchical headings
- We use <u>BM25F</u>
  - Scoring function for documents comprising multiple fields

$$\mathrm{BM25F} = \sum_{\kappa \in q} \frac{\mathrm{w}(\kappa, S)}{k_1 + \mathrm{w}(\kappa, S)} \log \frac{N - \mathrm{sf}(\kappa) + 0.5}{\mathrm{sf}(\kappa) + 0.5}, \quad \mathrm{w}(\kappa, S) = \frac{\mathrm{occurs}(\kappa, f, S) \cdot boost_f}{(1-b) + b \cdot \frac{\mathrm{length}(f, S)}{\mathrm{avgLength}(f)}}$$

Query Keywords in headings

- The sentence "Slow running."
  - NOT include a query keyword "jogging" in the sentence itself
  - Includes "jogging" in the hierarchical headings of the sentence



Popular exercise

Running

Jogging
Slow running. One benefit is to improve fitness.

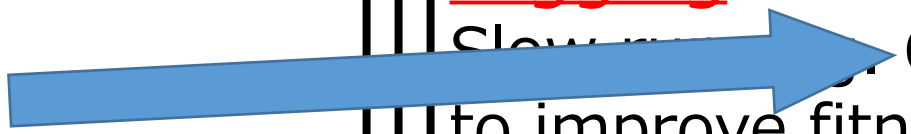
Sprint
Benefit of sprint is weight loss.

Swimming

Front Crawl
One benefit of this exercise is protection from stress.

Query Keywords in headings

- The sentence "One benefit is…"
  - includes a query keyword "jogging" in the hierarchical headings
  - Includes another query keyword "benefit" in the sentence itself
  - Important for a query "jogging benefit"



Popular exercise

Running

Jogging

Slow r… One benefit is to improve fitness.

Sprint

Benefit of sprint is weight loss.

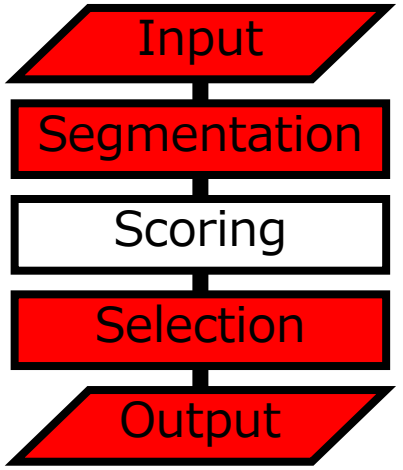
Swimming

Front Crawl

One benefit of this exercise is protection from stress.

3. Our Method

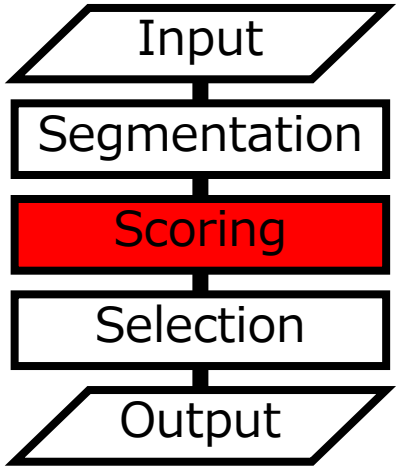
- Other steps are same as those of the existing method



Four Methods

1. Baseline method
   - Query keywords in sentence indicate importance
2. *Existing* method
   - Heading words in sentence also indicate importance
3. *Our* method
   - Query keywords in headings also indicate importance of sentences in their associated blocks
4. *Combined* method
   - All three ideas

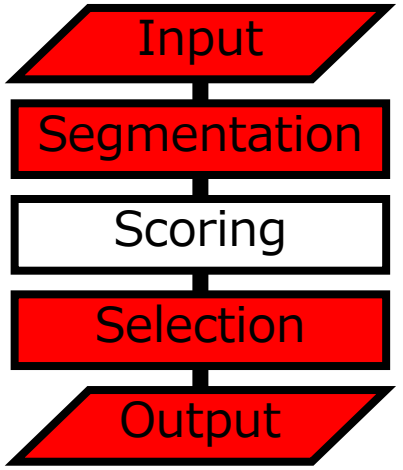
4. Combined Method



- The two ideas are independent
- The *combined* method adopts both
- The scoring function is:

$$\mathrm{BM25F}(\text{Query only}) + \mathrm{BM25F}(\text{Heading only}) + \mathrm{BM25F}(\text{Query heading})$$

4. Combined Method



- Other steps are same as
  those of the existing and our methods

Four Methods

1. Baseline method
   - Query keywords in sentence indicate importance
2. *Existing* method
   - Heading words in sentence also indicate importance
3. *Our* method
   - Query keywords in headings also indicate importance of sentences in their associated blocks
4. *Combined* method
   - All three ideas

Current Evaluation Result

On web search

Evaluation Methodology

- The most important feature of snippets: Judgeability
  - To what extent snippets help to judge relevance of pages
- In the INEX snippet retrieval track
  - Relevance judgments under different conditions are compared
    - Based on the entire documents
    - Only based on their snippets
  - If they agree, the snippets provides high judgeability and the snippet generation method is effective
  - Length limit of snippets: 180 letters for a page

Trappett, M., Geva, S., Trotman, A., Scholer, F., Sanderson, M.:
Overview of the INEX 2013 snippet retrieval track. In: CLEF (2013)

Data Set

- Target of INEX is XML while our target is web
- We used data set for TREC 2014 web track ad-hoc task
  - 50 keyword queries
  - ClueWeb12 document collection
  - Relevance judgement based on the entire pages
- We used only subset of the collection
  - Top-20 pages for each query (1,000 in total) from baseline
  - Generated by Indri with Waterloo spam filter

User Experiment

- To obtain snippet-based relevance judgment
- With 4 participants
- In each period, each participant is required to:
  1. Read intent description behind a query
  2. Scan titles and snippets of top-20 search result items
  3. Judge whether each page is relevant by only them
- Each snippets was judged once, each participant judged a page once and used all methods evenly

Evaluation Measures

- From INEX
  - Recall
    - $$\frac{|\text{Pages correctly judged as relevant on their snippets}|}{|\text{Pages relevant as a whole}|}$$
  - Negative recall
    - $$\frac{|\text{Pages correctly judged as irrelevant on their snippets}|}{|\text{Pages irrelevant as a whole}|}$$
  - Geometric mean of them
    - $$\sqrt{\text{Recall} \cdot \text{Negative Recall}}$$
    - Primary evaluation measure

General Queries

- Comparison of mean evaluation scores

Method	Recall	Negative recall	Geometric mean
Baseline	**0.475**	**0.828**	**0.512**
Existing	0.373	0.780	0.386
Ours	0.438	0.777	0.456
Combined	0.396	0.776	0.401

General Queries

- Comparison of mean evaluation scores

Method	Recall	Negative recall	Geometric mean
Baseline	**0.475**	**0.828**	**0.512**
Existing	0.373	0.780	0.386
Ours	0.438	0.777	0.456
Combined	0.396	0.776	0.401

Single Queries

- TREC splits queries into several types
- *Single* queries have clear and focused intents

Method	Recall	Negative recall	Geometric mean
Baseline	0.431	**0.837**	0.488
Existing	0.336	0.804	0.392
Ours	0.375	0.816	0.443
Combined	**0.491**	0.795	**0.530**

Single Queries

- TREC splits queries into several types
- *Single* queries have clear and focused intents

Method	Recall	Negative recall	Geometric mean
Baseline	0.431	**0.837**	0.488
Existing	0.336	0.804	0.392
Ours	0.375	0.816	0.443
Combined	**0.491**	0.795	**0.530**

Query Length and Geometric Mean Scores

- Users need pages including all keywords in relation
  - Finding such pages gets more difficult for longer queries
  - Headings may help indicating relation

Method	2 keywords	3 keywords	4+ keywords
Baseline	**0.585**	**0.503**	0.394
Existing	0.406	0.387	0.350
Ours	0.543	0.378	0.362
Combined	0.393	0.388	**0.425**

Query Length and Geometric Mean Scores

- Users need pages including all keywords in relation
  - Finding such pages gets more difficult for longer queries
  - Headings may help indicating relation

Method	2 keywords	3 keywords	4+ keywords
Baseline	**0.585**	**0.503**	0.394
Existing	0.406	0.387	0.350
Ours	0.543	0.378	0.362
Combined	0.393	0.388	**0.425**

Conclusion

- Introduced a new idea for heading-aware snippet generation
  - Query keyword in headings of sentences indicate importance of them
- Compared baseline and 3 heading-aware generation methods
- Evaluation result indicated that heading-aware methods were:
  - Not effective for general queries
  - Effective only for queries:
    - representing its intent clearly
    - containing four or more keywords
- Additional evaluation with more queries is needed