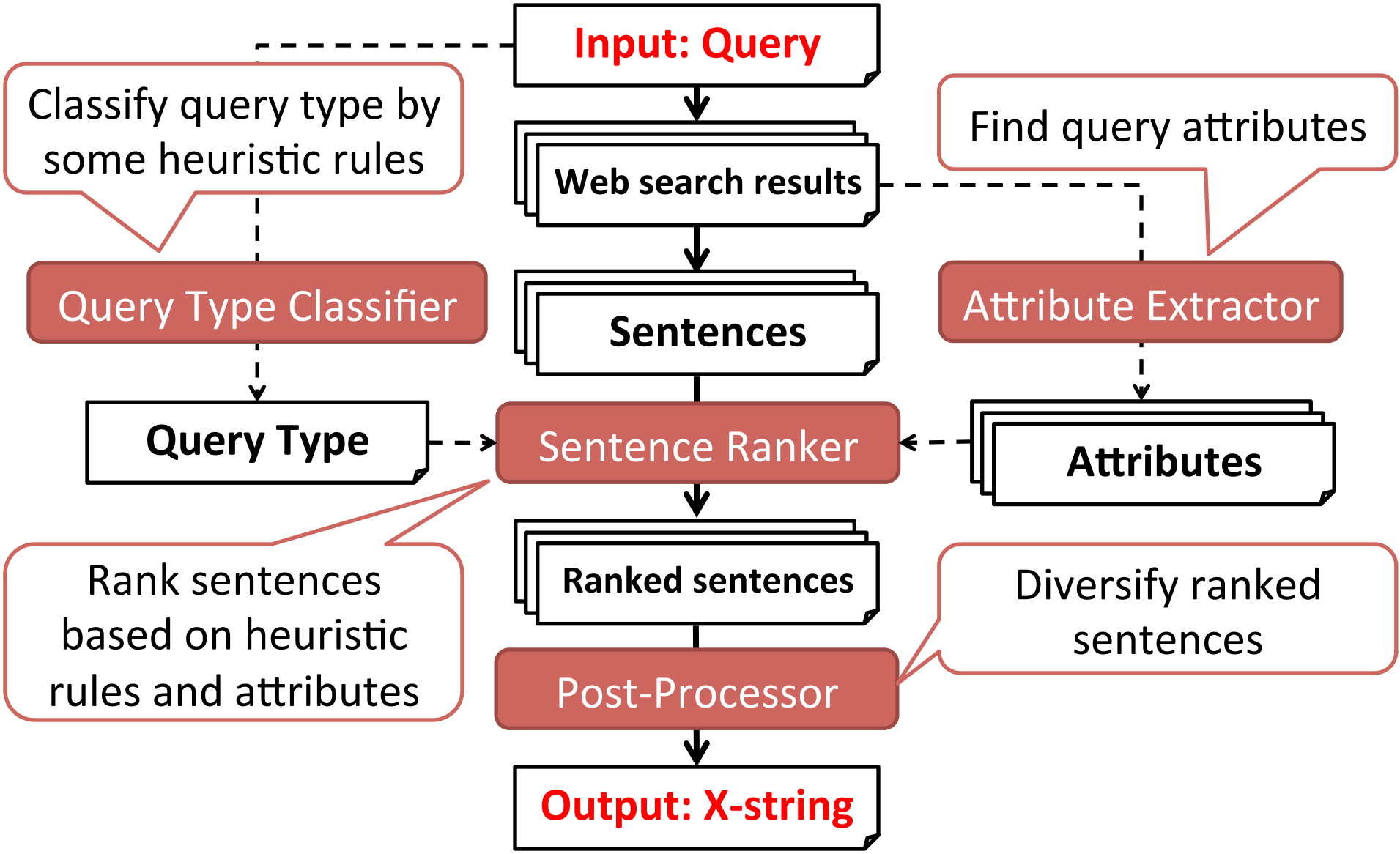
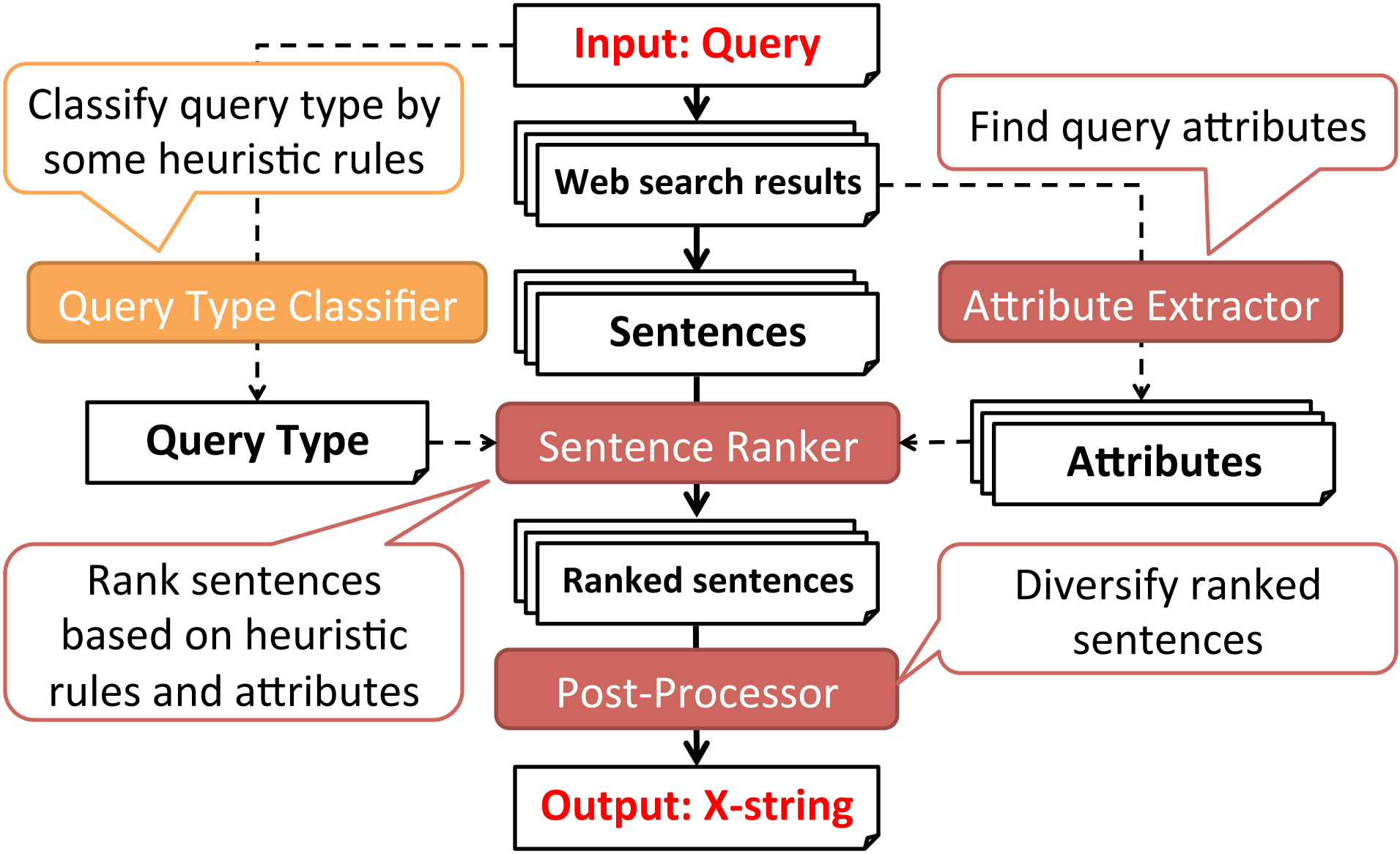
# MSRA at NTCIR-10 1CLICK-2

Kazuya Narita (*Tohoku University, Japan*),
Tetsuya Sakai, Zhicheng Dou (*MSRA, China*),
Young-In Song (*NHN Corporation, Korea*)

# Overview


Input: Query
Classify query type by some heuristic rules
Find query attributes
Web search results
Query Type Classifier
Sentences
Attribute Extractor
Query Type
Sentence Ranker
Attributes
Rank sentences based on heuristic rules and attributes
Ranked sentences
Diversify ranked sentences
Post-Processor
Output: X-string

# Overview


Input: Query
Classify query type by some heuristic rules
Find query attributes
Web search results
Query Type Classifier
Sentences
Attribute Extractor
Query Type
Sentence Ranker
Attributes
Rank sentences based on heuristic rules and attributes
Ranked sentences
Diversify ranked sentences
Post-Processor
Output: X-string

# Classification Rules

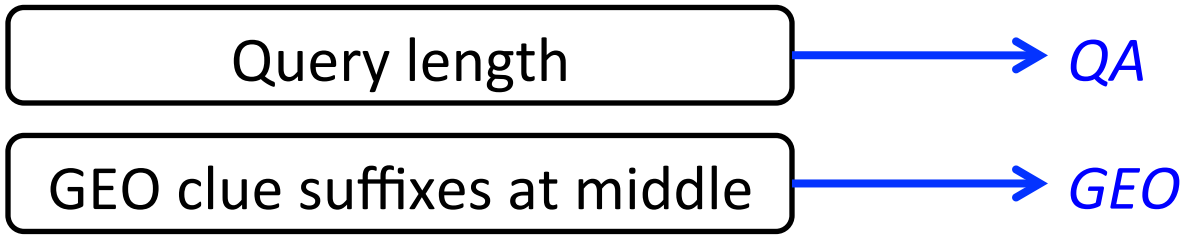
- Classification is based on some heuristic rules:
  - E.g. Query length → *QA*

"世界で初めてノーベル賞をとったのは誰か"
(*Who took the Nobel Prize for the first time in the world?*)

Long query may be QA

# Classification Rules

- Classification is based on some heuristic rules:
  - E.g.

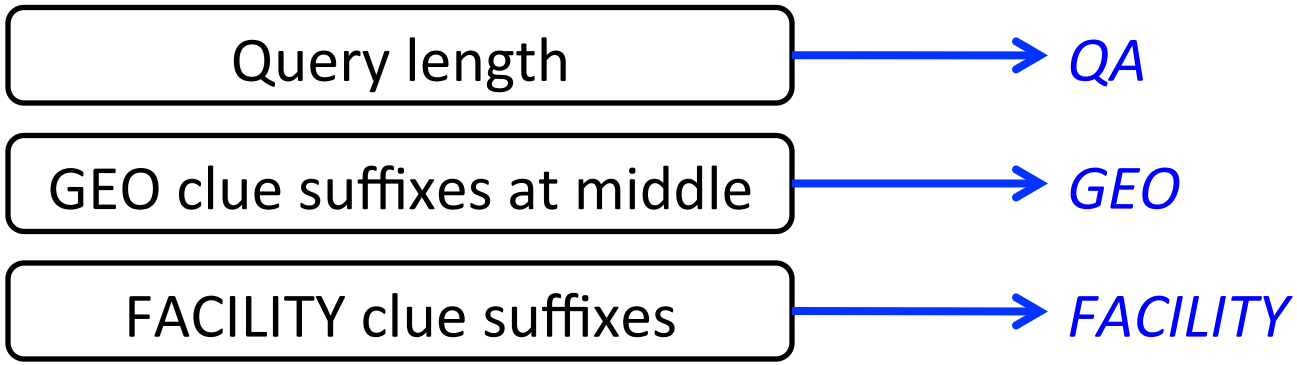



"博多駅 ホテル"
(*hotel near Hakata station*)

Other suffixes: "市" (*city*), "区" (*ward*), "町" (*town*) ...

# Classification Rules

- Classification is based on some heuristic rules:
  - E.g.

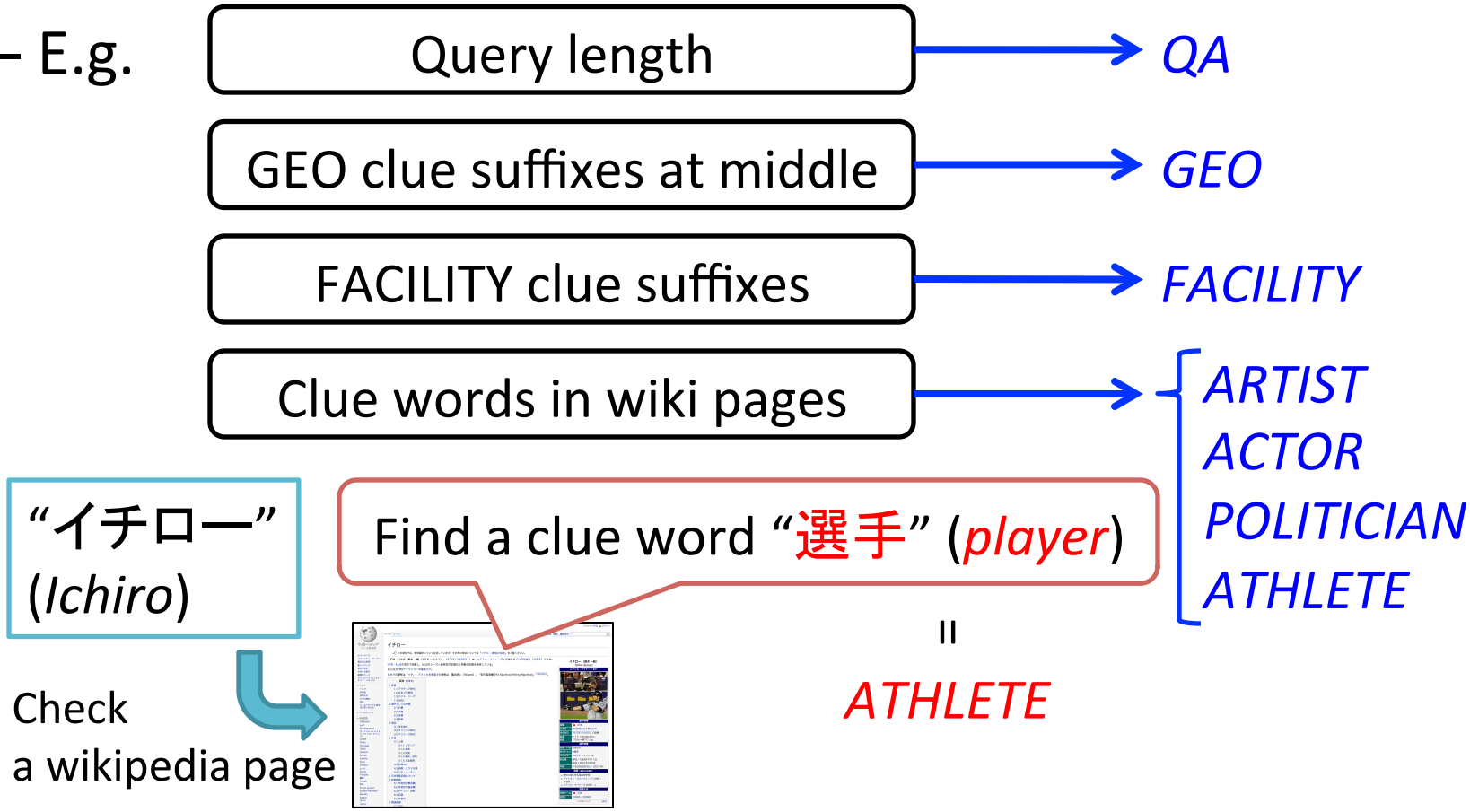



"須磨海浜水族園"
(*Suma Aqualife Park*)

Other suffixes: "学校" (*school*), "病院" (*hospital*) ...

# Classification Rules

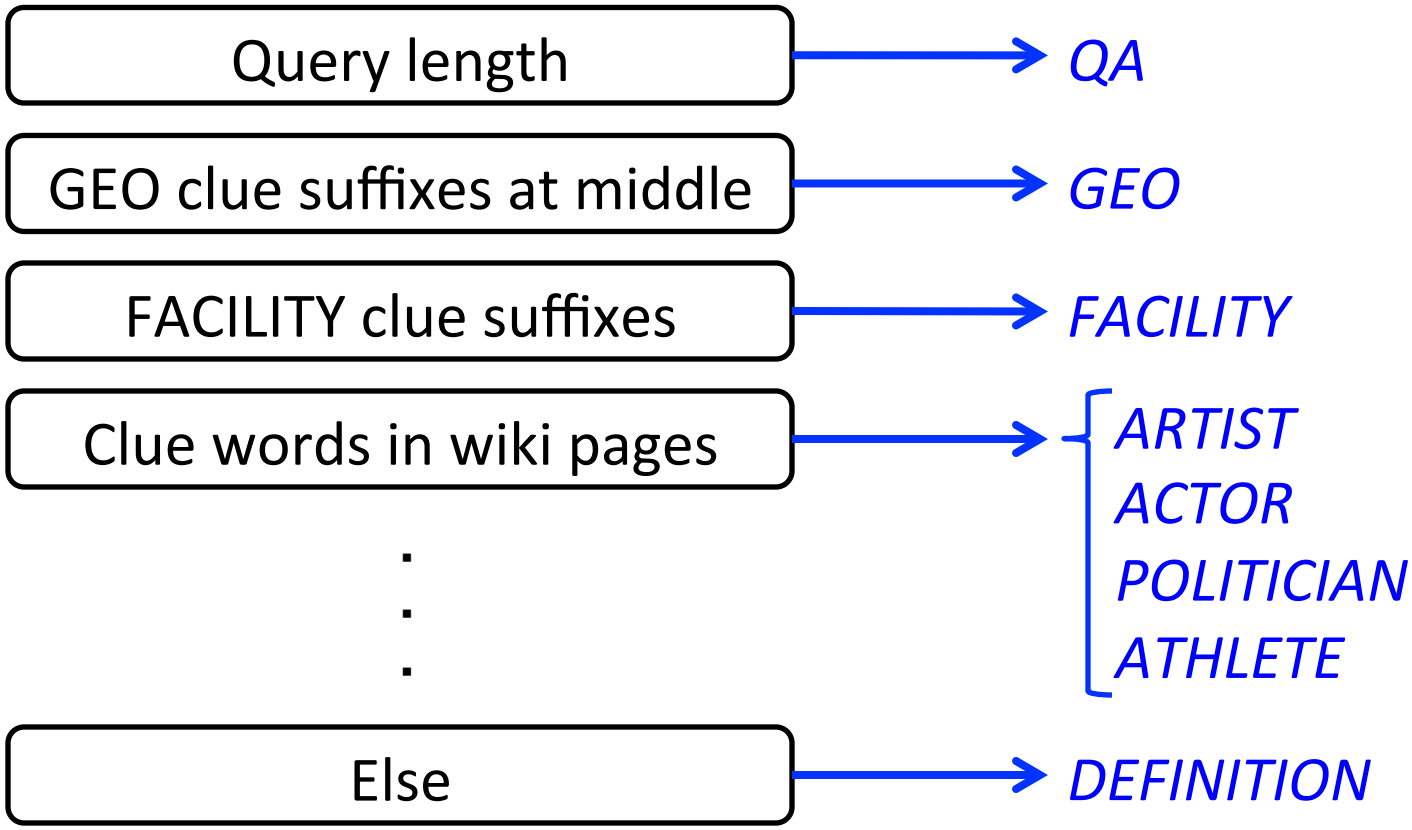
- Classification is based on some heuristic rules:
  - E.g.




Other clue words: "小説家" (*novelist*), "女優" (*actress*) ...

# Classification Rules

- Classification is based on some heuristic rules:
  - E.g.



# Results of Query Type Classifier

- Accuracy: 83% (83/100)

Not good…

| sys \ gold | ART | ACT | POL | ATH | FAC | GEO | DEF | QA |
|---|---|---|---|---|---|---|---|---|
| ARTIST | 9 | 1 | 1 | | 1 | 1 | | |
| ACTOR | 1 | 9 | | | | | | |
| POLITICIAN | | | 8 | | | | | |
| ATHLETE | | | | 10 | | | 1 | |
| FACILITY | | | | | 7 | | 1 | |
| GEO | | | | | 1 | 14 | 1 | |
| DEFINITION | | | 1 | | 5 | | 12 | 1 |
| QA | | | | | 1 | | | 14 |

# Error Analysis for FACILITY

## Correct Example

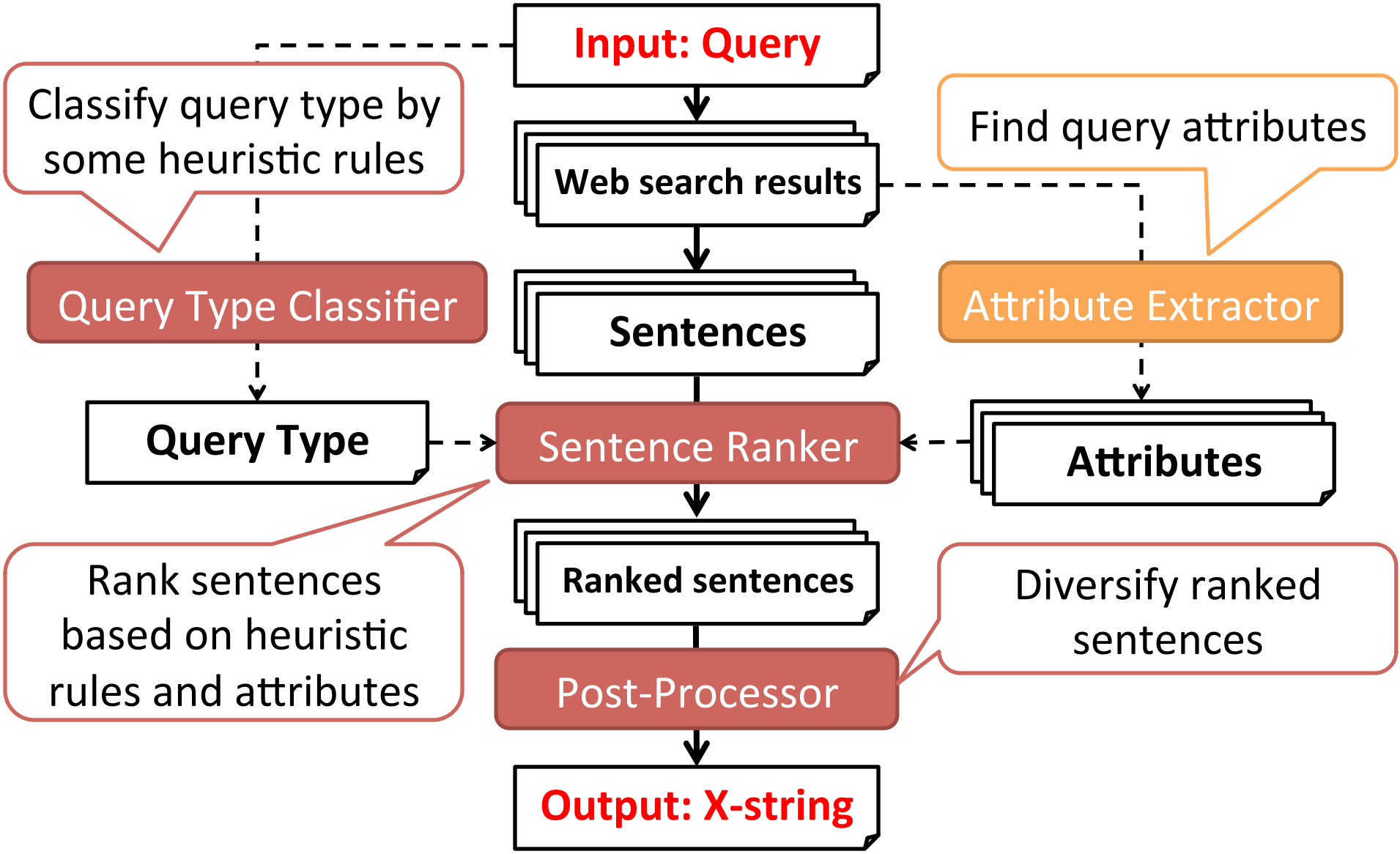
- Correct only by suffix rules
  - 須磨海浜水族園
    (*Suma Aqualife Park*)
  - 横浜市役所
    (*Yokohama City Hall*)
  - 小金井図書館
    (*Koganei library*)
  - ハワイパシフィック大学
    (*Hawaii Pacific University*)
  - あおやま矯正歯科医院
    (*Aoyama Orthodontic Office*)

## Wrong Example

- **Insufficient suffixes**
  - 京都真如堂
    (*Kyoto Shinnyo-do temple*)
  - 南ヶ丘牧場
    (*Minamigaoka dairy*)

- **Difficult only by suffix rules**
  - カーサ・ディ・ナポリ
    (*Casa di Napoli*)
  - ザ・ペニンシュラ東京
    (*The Peninsula Tokyo*)
  - らーめんてつや
    (*Ramen Tetsuya*)

# Overview


Input: Query
Classify query type by some heuristic rules
Find query attributes
Web search results
Query Type Classifier
Attribute Extractor
Sentences
Query Type
Sentence Ranker
Attributes
Rank sentences based on heuristic rules and attributes
Ranked sentences
Diversify ranked sentences
Post-Processor
Output: X-string

# Query Attributes

***Ichiro***


- ***Born***
- ***Height***
- ***Position***

***Future Science Museum***

- ***Address***
- ***Tel***
- ***Opening Hours***

# Query Attributes

***Ichiro***


***Future Science Museum***


These attributes may be useful when users are seeking for basic information of entities

***Born***

***Height***

***Position***

***Address***

***Tel***

***Opening Hours***

# Attribute Extraction


Web search results

# Attribute Extraction


Web search results

Extract tables or text containing ":"

| | |
|---|---|
| Born | October 22, 1973 (age 38) |
| Height | 5 ft 11 in (1.80m) |
| Position | Outfielder |

Born: October 22, 1973 (age 38)
Height: 5 ft 11 in (1.80m)
Position: Outfielder

# Attribute Extraction


Web search results

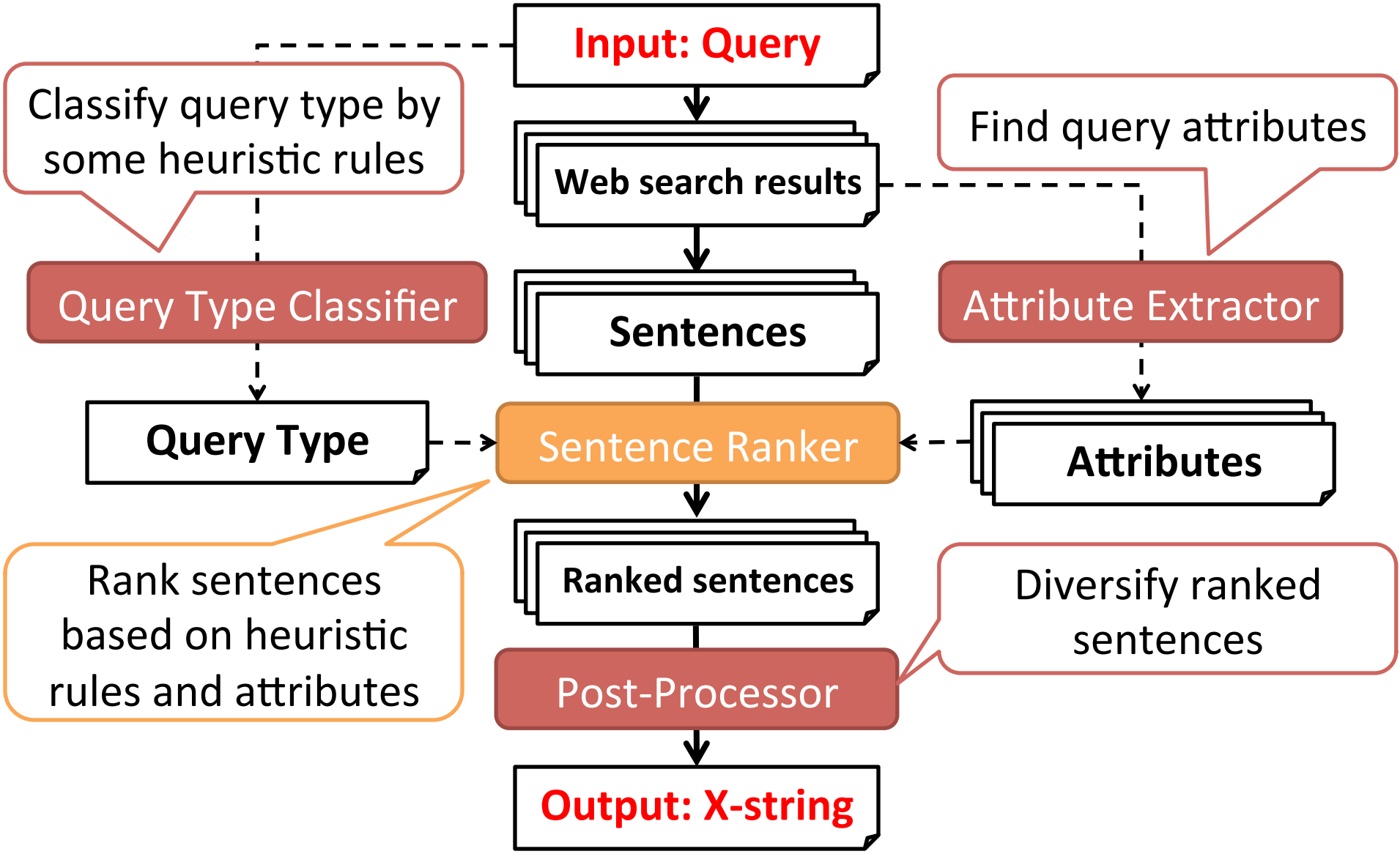
Extract as query attributes

Extract tables or text containing ":"

| | |
|---|---|
| ***Born*** | October 22, 1973 (age 38) |
| ***Height*** | 5 ft 11 in (1.80m) |
| ***Position*** | Outfielder |

***Born***: October 22, 1973 (age 38)
***Height***: 5 ft 11 in (1.80m)
***Position***: Outfielder

# Overview

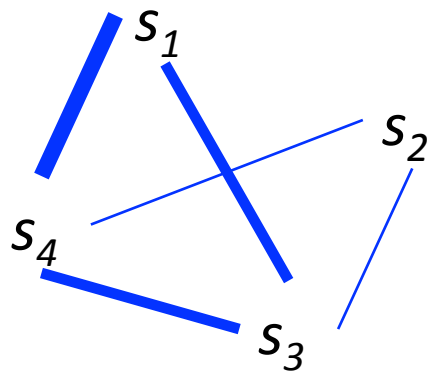

Input: Query
Classify query type by some heuristic rules
Find query attributes
Web search results
Query Type Classifier
Sentences
Attribute Extractor
Query Type
Sentence Ranker
Attributes
Rank sentences based on heuristic rules and attributes
Ranked sentences
Diversify ranked sentences
Post-Processor
Output: X-string

# Sentence Ranking

- Scoring based on following components:
  1. Content similarity
  2. Query attributes
  3. Manual clue words and URLs
  4. Search Rank

# Scoring

1. Content similarity (***LexRank*** [Erkan and Radev 2004])
   – Similar sentences are more relevant




Importance of each sentences
(similar to many other sentences -> High)

$$\mathbf{p} = \left[ d\mathbf{U} + (1-d)\mathbf{B} \right]^T \mathbf{p}$$

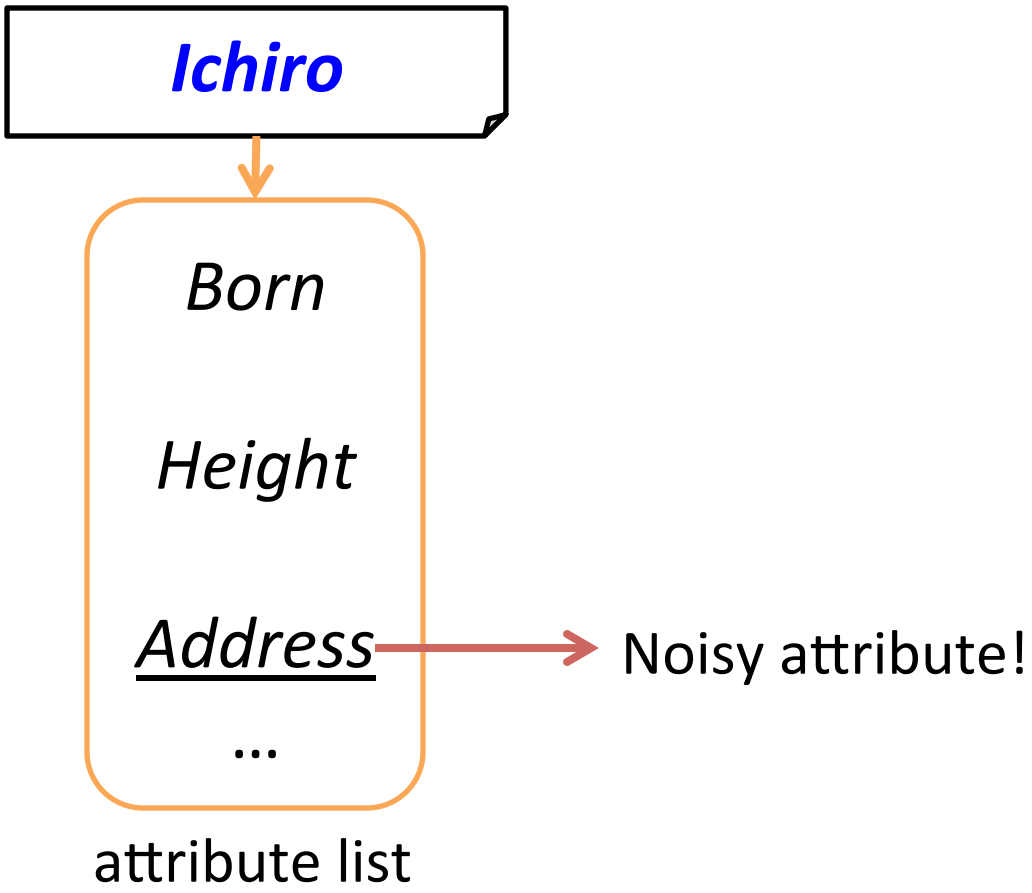
All elements are $1/n$
($n$: the number of sentences)

Adjacency matrix
of the cosine similarity

Damping factor

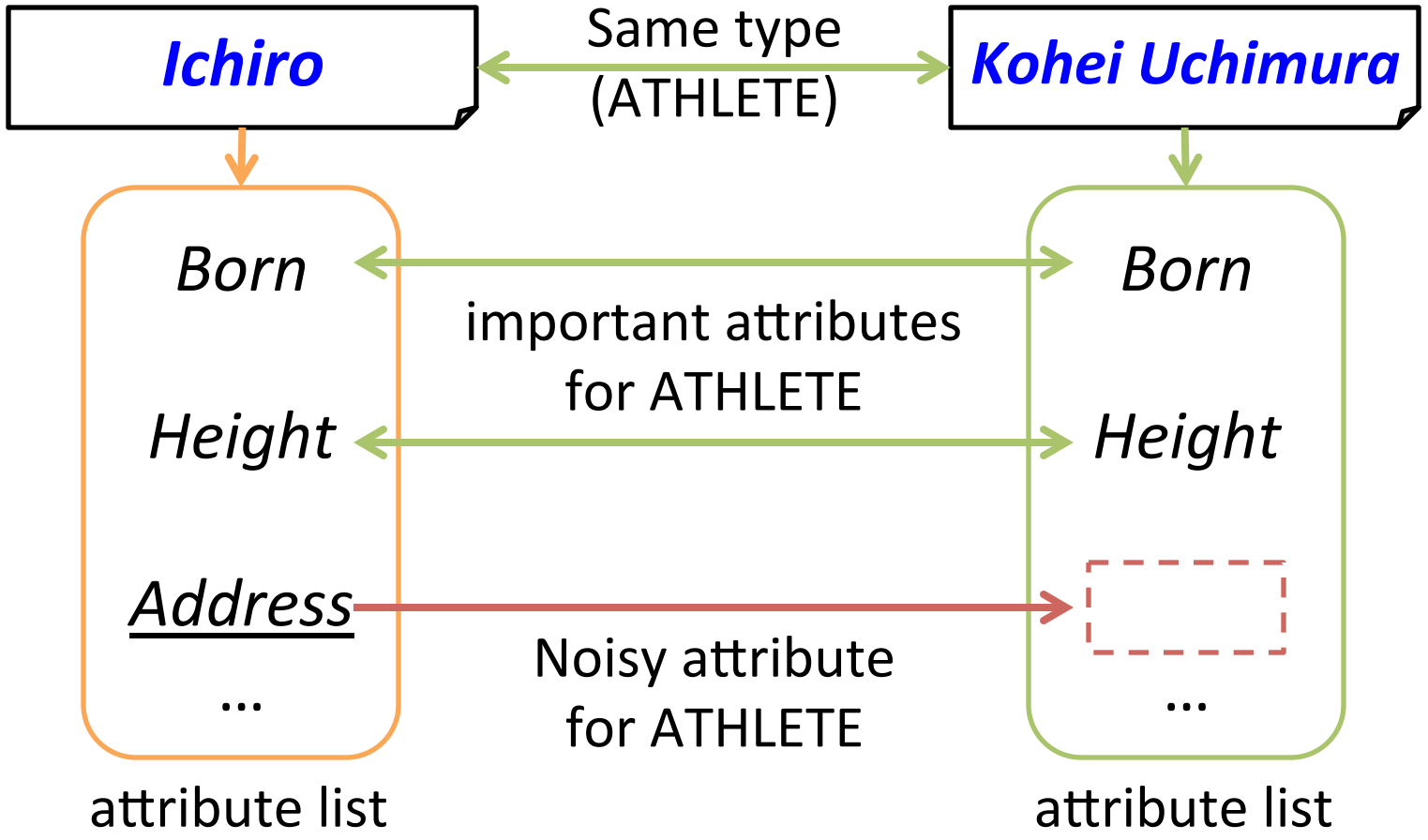
# Scoring

2. Query attributes
   – Attributes extracted automatically are *noisy*



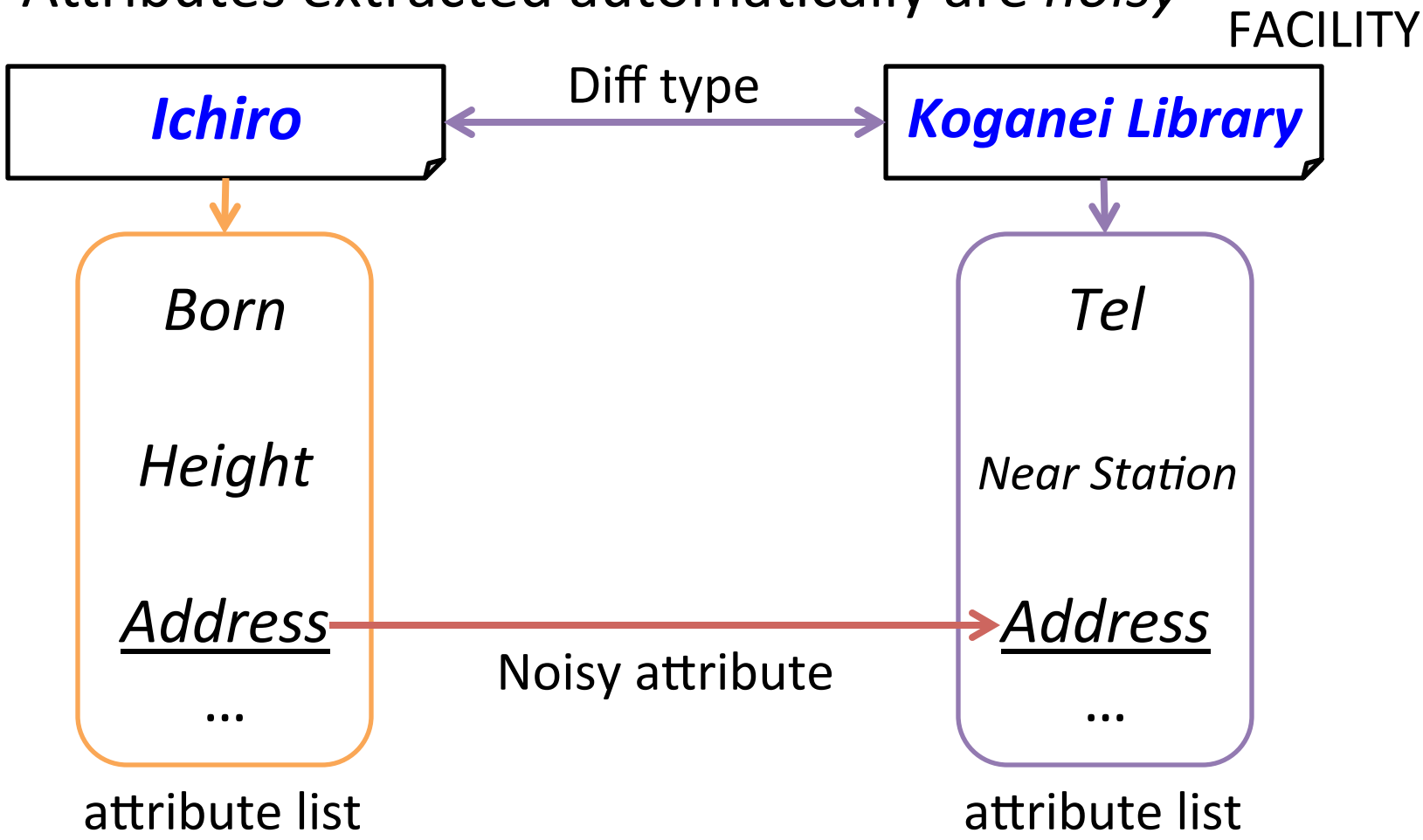
# Scoring

2. Query attributes
   – Attributes extracted automatically are *noisy*



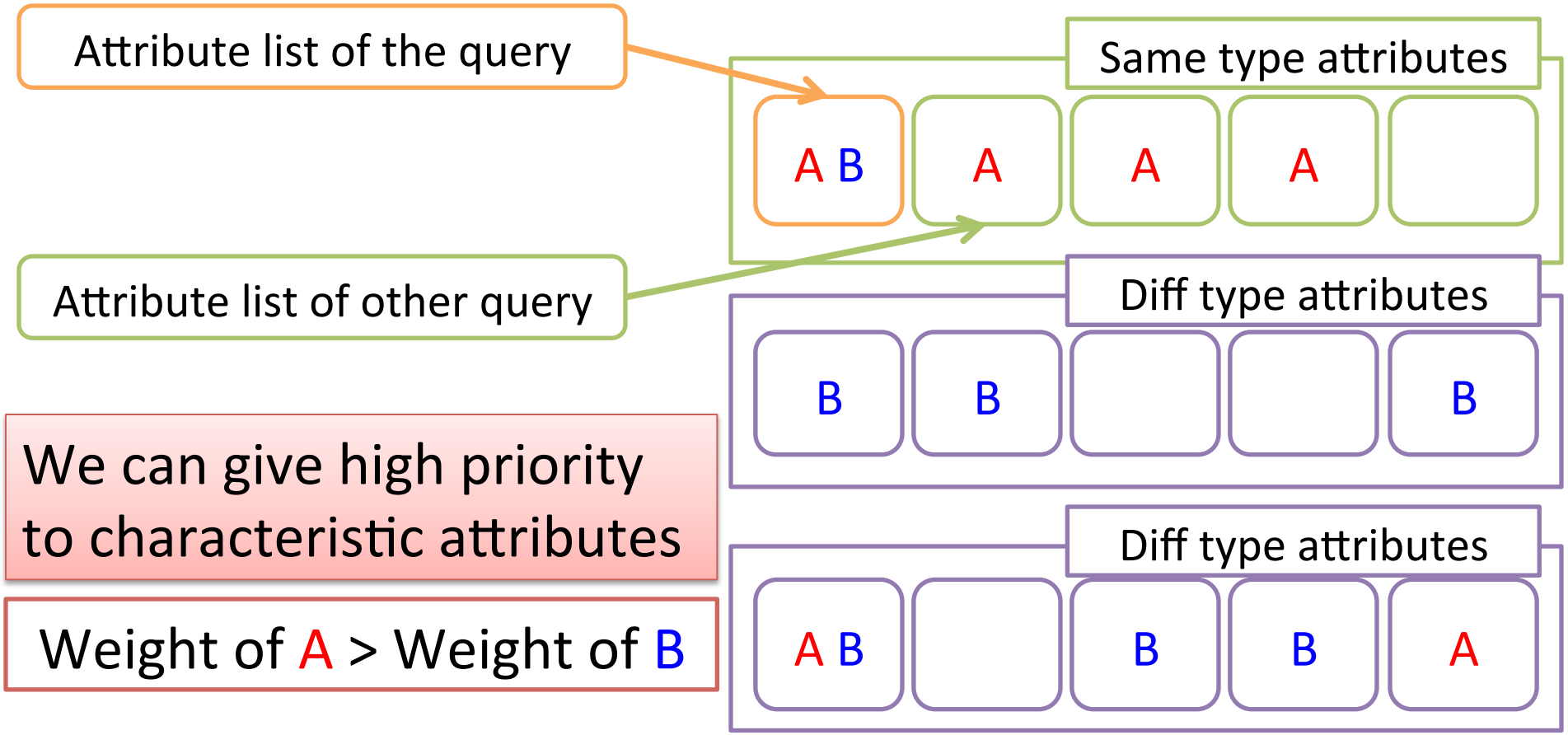
# Scoring

2. Query attributes
   – Attributes extracted automatically are *noisy*



# Scoring

2. Query attributes
   – Weight each attribute as importance like TF-IDF



# Scoring

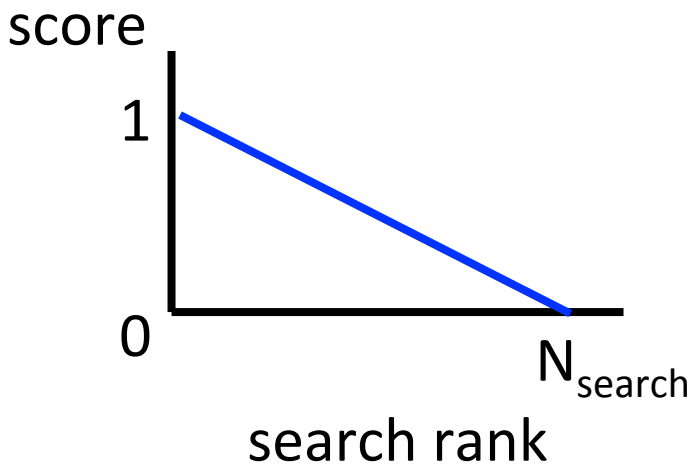
3. Manual clue words and URLs
   - "***birthday***" or "***birth town***": more relevant for celebrity
   - "***talent.yahoo.co.jp***": more relevant for celebrity
   - "***address***" or "***phone number***": more relevant for facility

   - "***posted by***" or "***answerer***": less relevant for all
   - "***amazon.co.jp***": less important for all

# Scoring

4. Search rank
   – Sentences of high rank webpage is more important

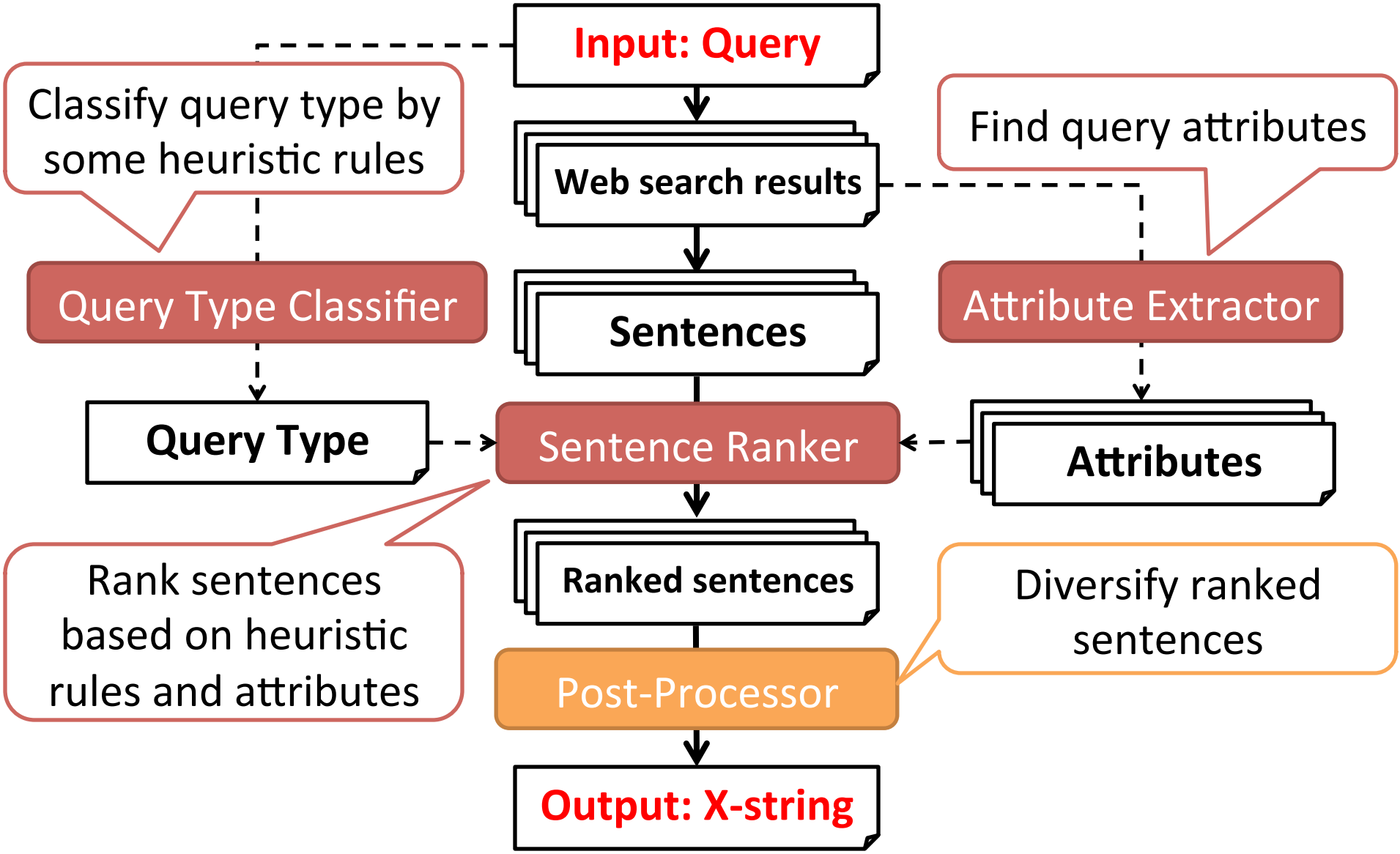
$$S_{SRank}(s) = 1 - \frac{rank(s)}{N_{search}}$$



# Sentence Ranking

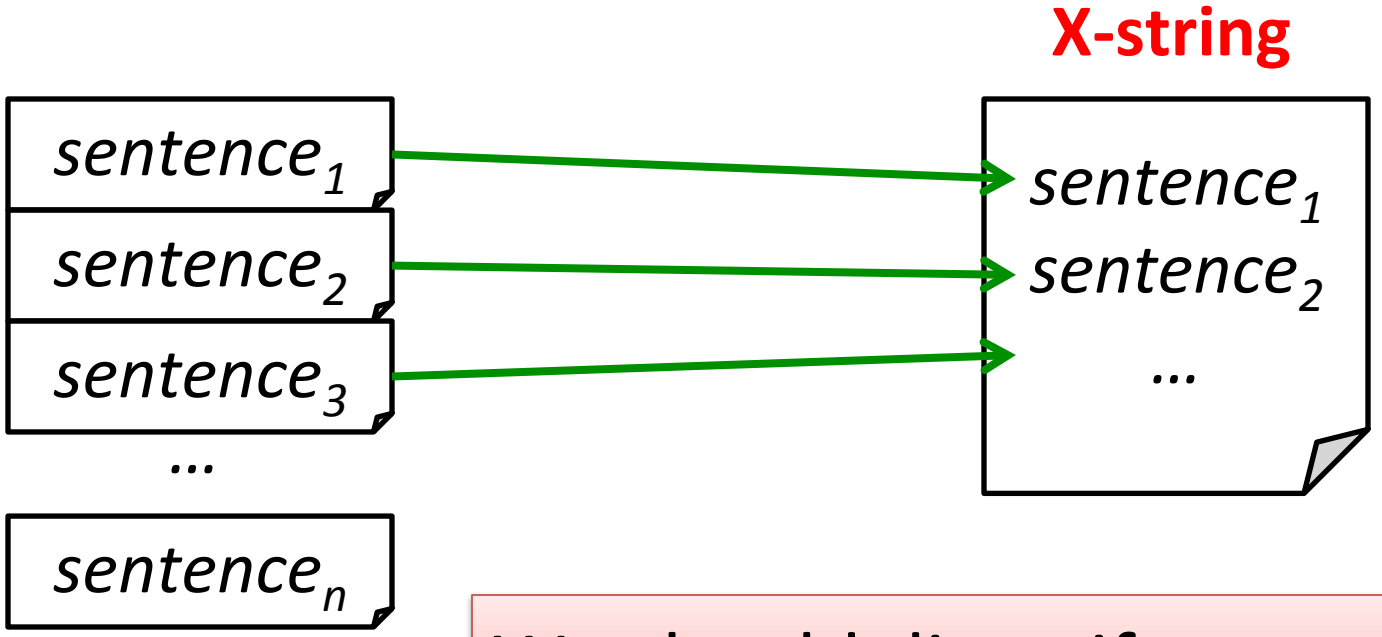
- Scoring based on following components:
  1. Content similarity
     - Similar sentences are more relevant
  2. Query attributes
     - Weight for each attribute as importance like TF-IDF
  3. Manual clue words and URLs
  4. Search Rank
     - Sentences of high rank webpage is more important

# Overview


Input: Query
Classify query type by some heuristic rules
Find query attributes
Web search results
Query Type Classifier
Sentences
Attribute Extractor
Query Type
Sentence Ranker
Attributes
Rank sentences based on heuristic rules and attributes
Ranked sentences
Diversify ranked sentences
Post-Processor
Output: X-string

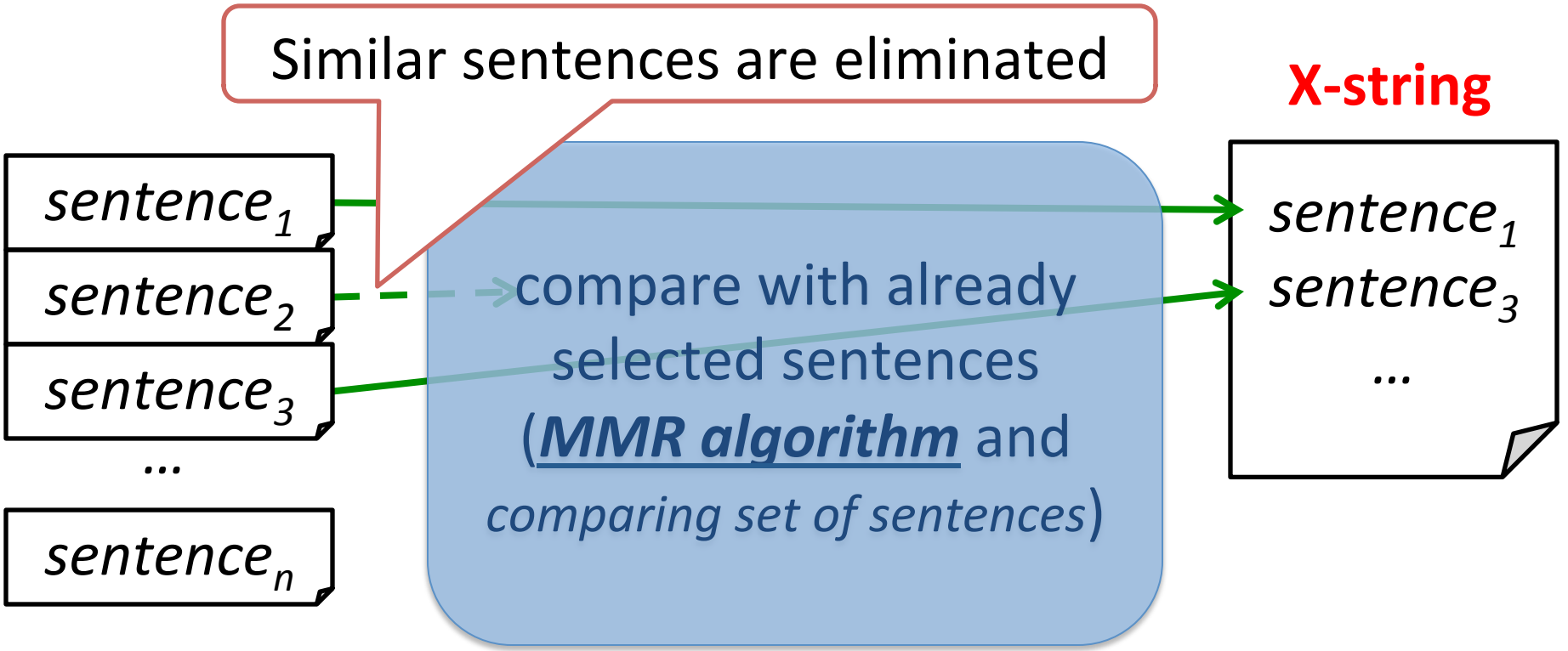
# Sentence diversity

- Output top *N* sentences as X-string
  - But similar sentences tend to be output
    - Redundancy is undesirable for 1CLICK




We should diversify sentences

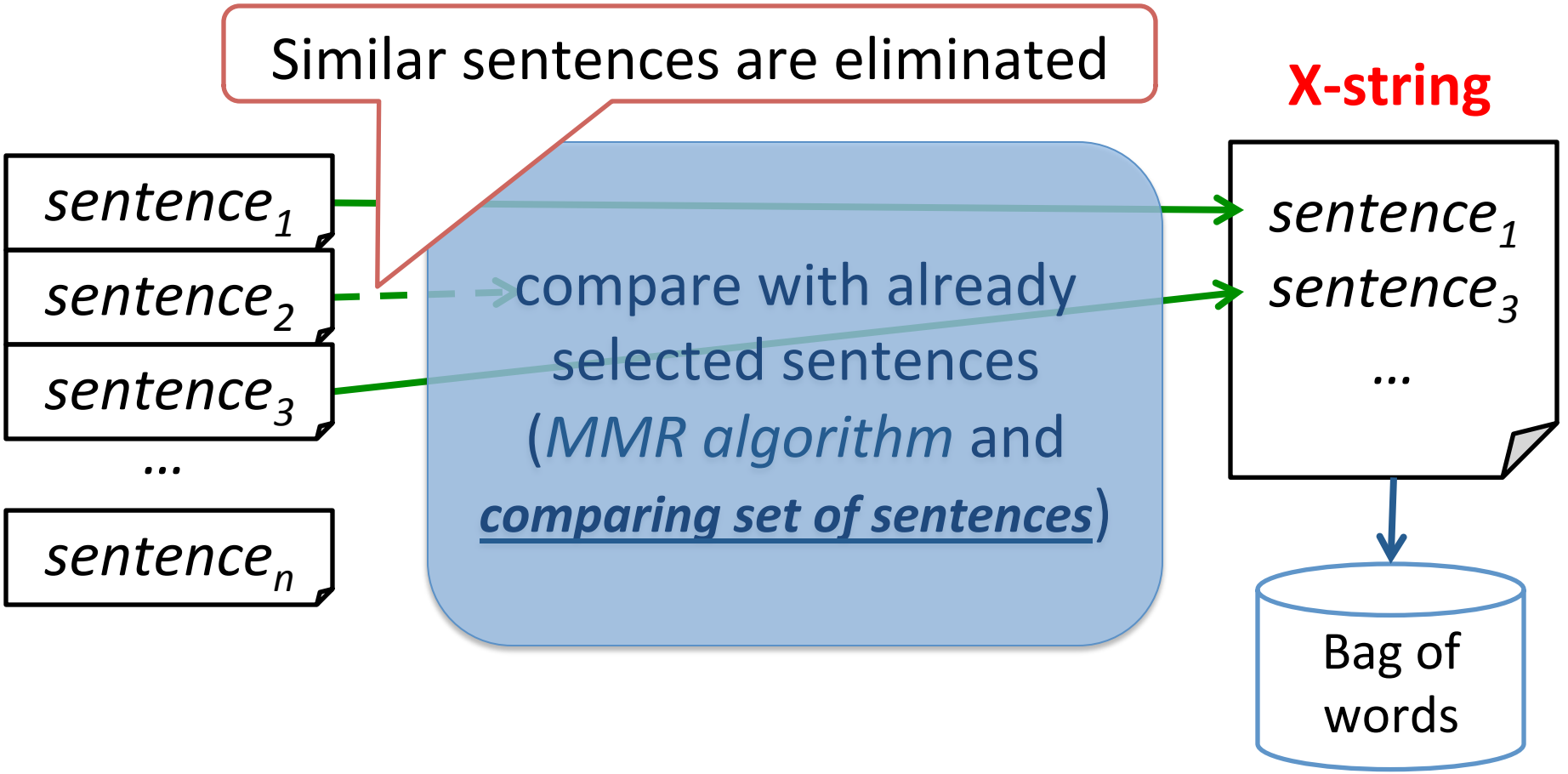
# To consider sentence diversity




$$MMR = \arg\max_{s_i \in R \setminus S} \left[ \lambda Score(s_i) - (1-\lambda) \max_{s_j \in S} Sim(s_i, s_j) \right]$$

Give a penalty depending on similarity to already selected sentences

# To consider sentence diversity

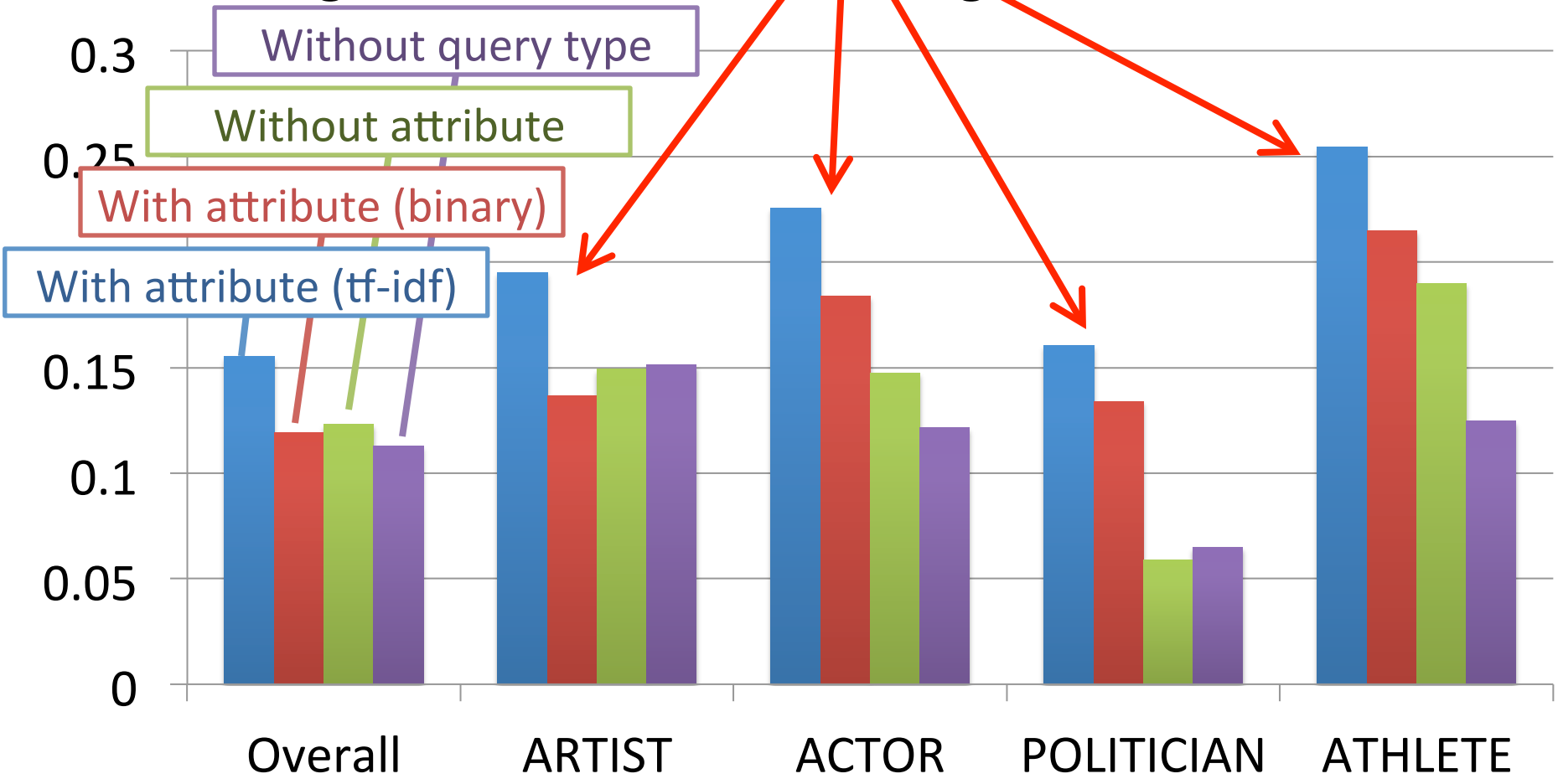



Compare a sentence with bag of words of already selected sentences

# EVALUATION RESULTS

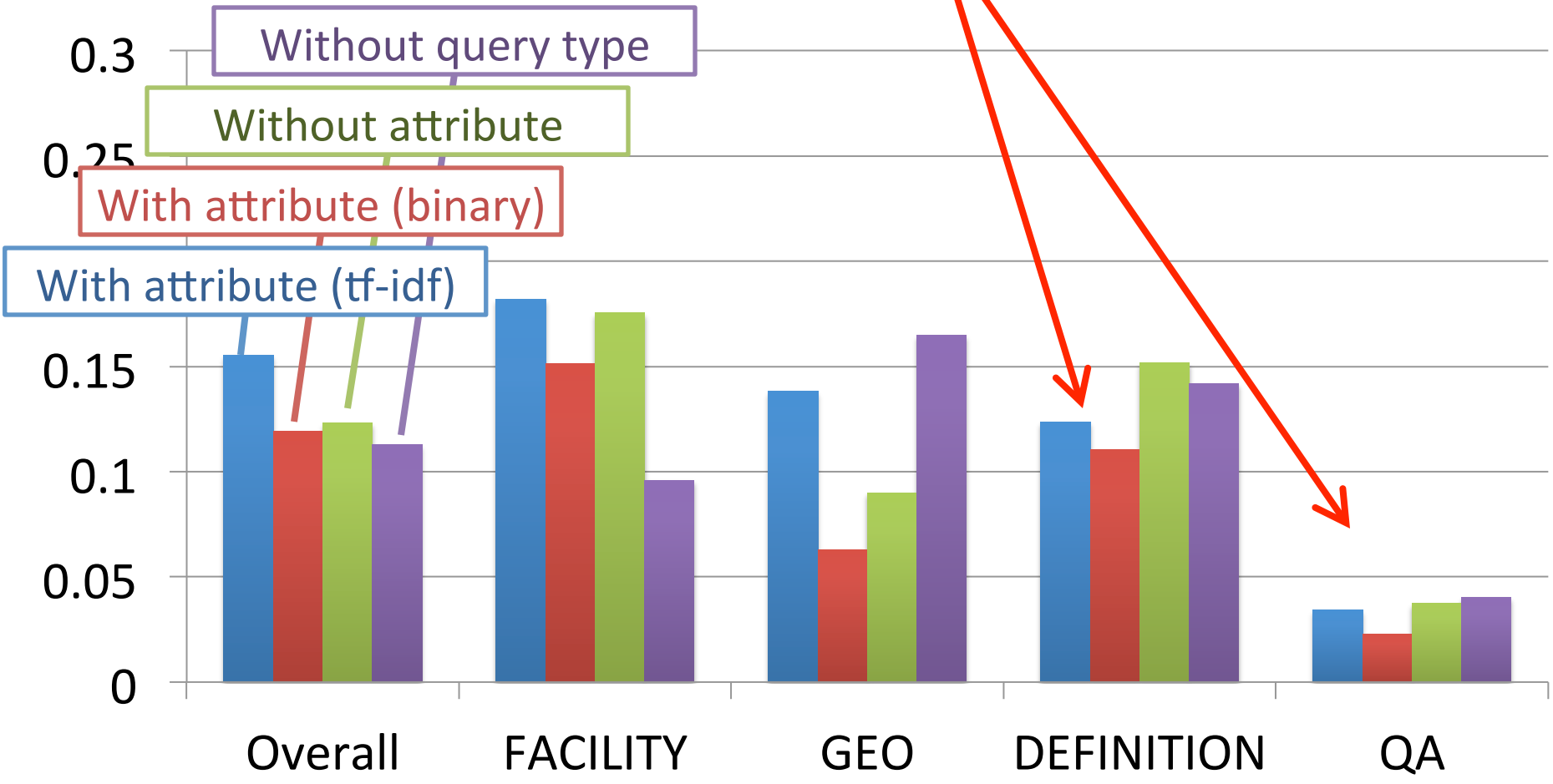
# Results (mean S#-measures)

- Query attributes are effective for celebrity
  - Though the difference is not significant



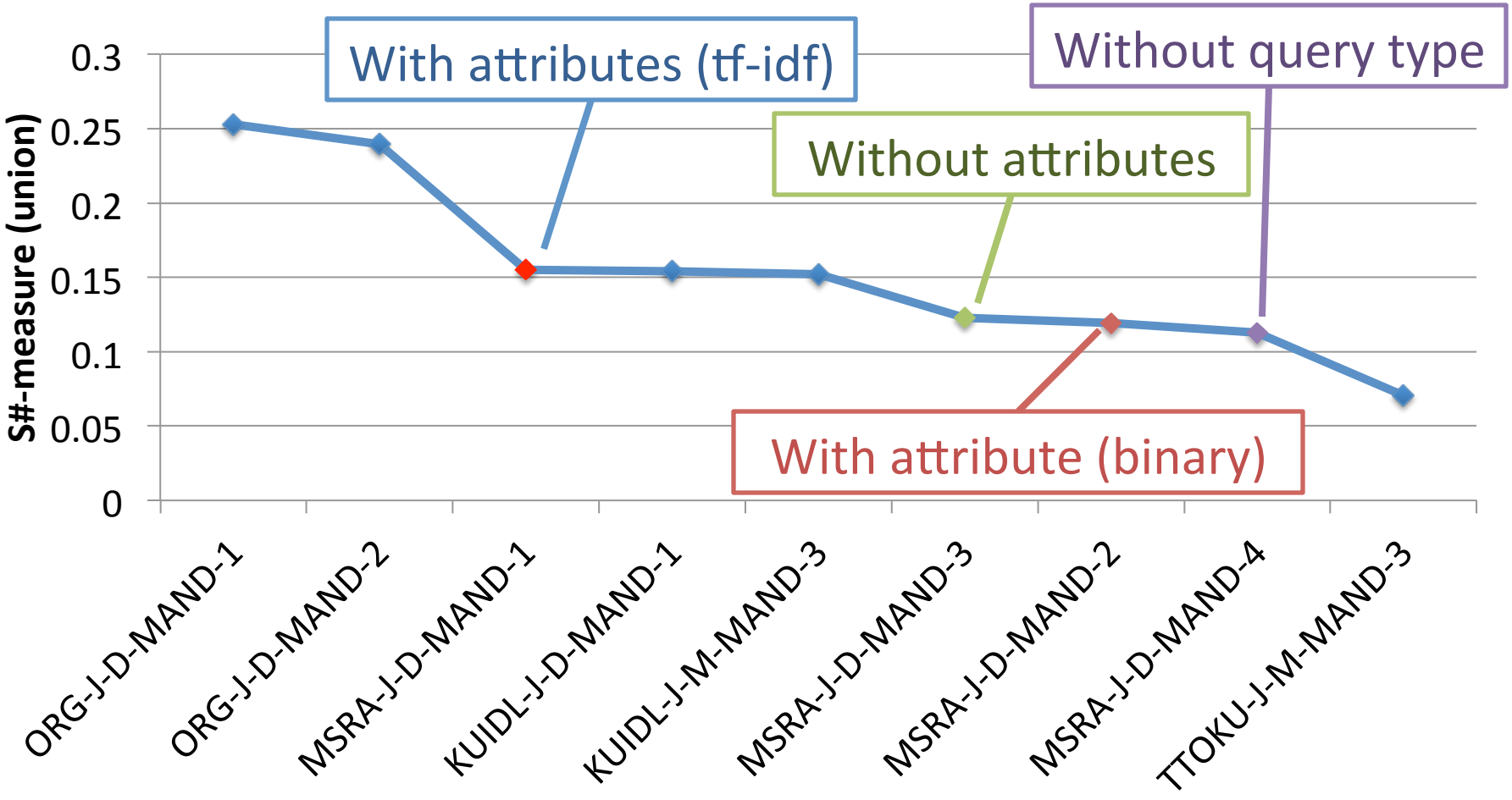
# Results (mean S#-measures)

- Query attributes do not work well for DEFINITION and QA queries



# Comparison with other participants

- Top performance among MANDATORY runs
  - Though the difference is not significant



# Conclusions

- Automatic query attribute extraction is effective especially for celebrity queries
  - tf-idf weight is effective
- Future work:
  - Automatic extraction of clue words for query type classification
  - New framework instead of query attributes for DEFINITION and QA queries