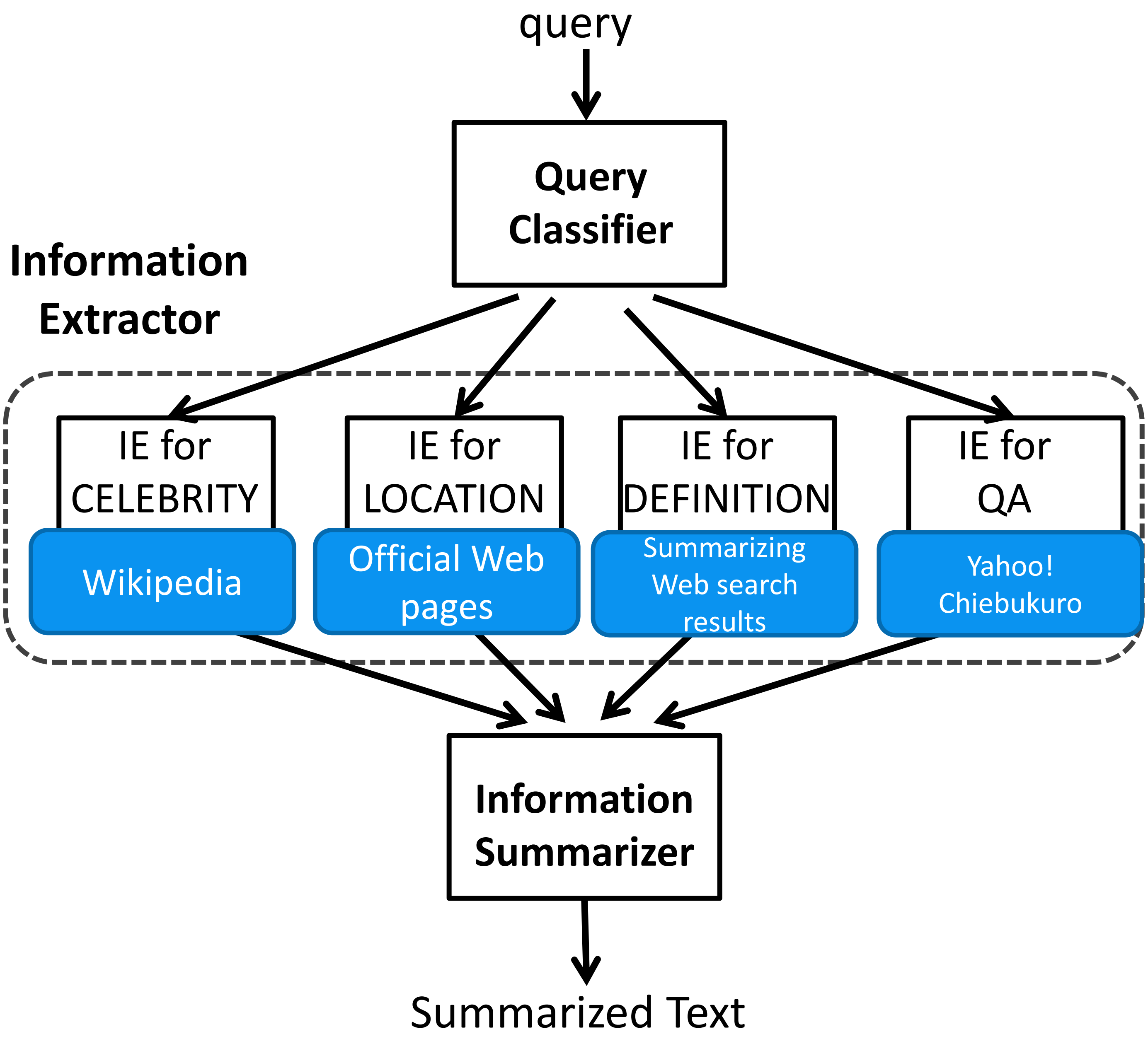
Information Extraction based Approach for the NTCIR-9 1CLICK Task

Meng Zhao, Kosetsu Tsukuda, Yoshiyuki Shoji, Makoto P. Kato,
Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka
Graduate School of Informatics, Kyoto University

NTCIR-9 Workshop Meeting, 2011/12/9, NII Tokyo, Japan

Our Framework



Query Classifier

Classify four types of query by multi class SVM

185 Features

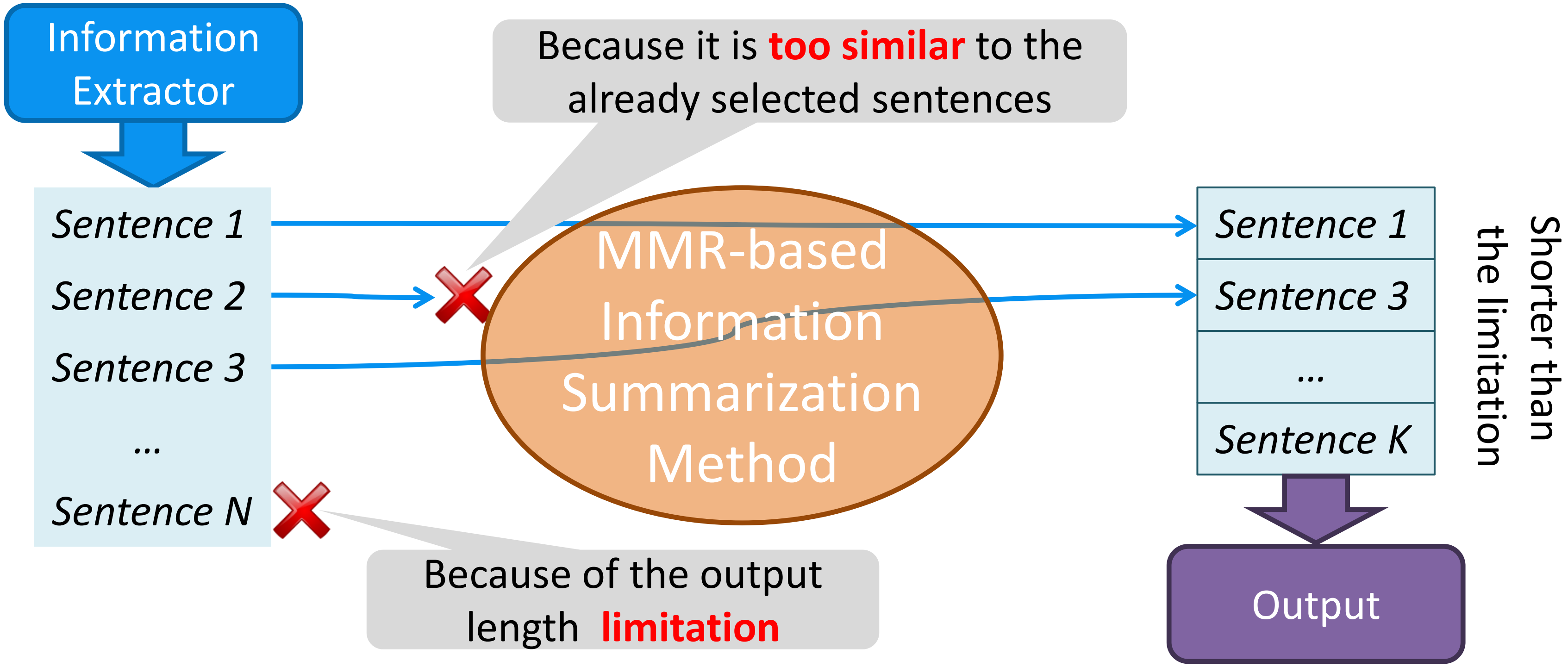
Feature	**# of features**
Has Wikipedia article	1
Frequency of parts-of-speech	44
Query unigram	85
Sentence pattern	2
Number of documents containing expanded query	15
Has travel services	1
Number of search results	1
Terms in search results	39
Total	**185**

Search with 15 expanded query such as “query-san(さん)” and “query-senshu(選手)” to distinguish name of CELEBRITY.

Count selected 39 terms such as “profile,” “born at” and “Chome (丁目)” in snippets of Yahoo Japan search result

Information Summarizer

- Eliminates overlap from the extracted sentences.



Information Extractor

Sentence 1
Sentence 2
Sentence 3
…
Sentence N

Because it is **too similar** to the already selected sentences

MMR-based Information Summarization Method

Because of the output length **limitation**

Sentence 1
Sentence 3
…
Sentence K

Shorter than the limitation

Output

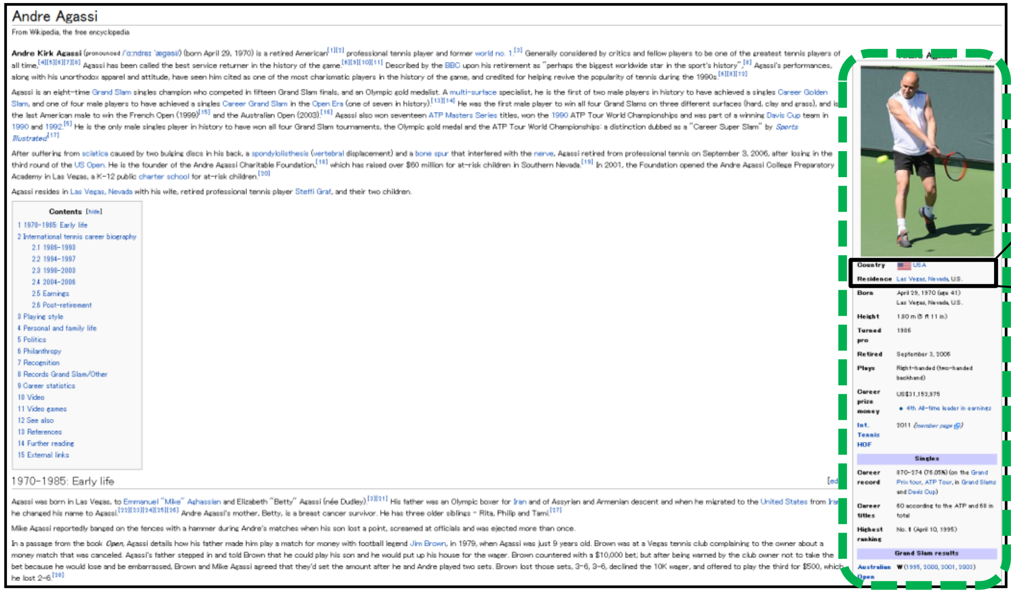
$$MMR = \underset{d_i \in D/S}{\operatorname{argmax}} [\lambda(\boxed{Score(d_i)}) - (1-\lambda)\boxed{\underset{d_j \in S}{\operatorname{maxsim}}(d_i, d_j)}]$$

Score of the sentence

Similarity to the output sentences

IE for CELEBLITY (Attribute-based)

- Extract pairs of an attribute name and its value from *Infobox* in Wikipedia



Andre Agassi

From Wikipedia, the free encyclopedia

Andre Kirk Agassi (pronounced /ˈɑːndreɪ ˈæɡəsi/) (born April 29, 1970) is a retired American[1][2] professional tennis player and former world no. 1.[3] Generally considered by critics and fellow players to be one of the greatest tennis players of all time,[4][5][6][7][8] Agassi has been called the best service returner in the history of the game.[6][9][10][11] Described by the BBC upon his retirement as "perhaps the biggest worldwide star in the sport's history",[8] Agassi's performances, along with his unorthodox apparel and attitude, have seen him cited as one of the most charismatic players in the history of the game, and credited for helping revive the popularity of tennis during the 1990s.[6][8][12]

Agassi is an eight-time Grand Slam singles champion who competed in fifteen Grand Slam finals, and an Olympic gold medalist. A multi-surface specialist, he is the first of two male players in history to have achieved a singles Career Golden Slam, and one of four male players to have achieved a singles Career Grand Slam in the Open Era (one of seven in history).[13][14] He was the first male player to win all four Grand Slams on three different surfaces (hard, clay and grass), and is the last American male to win the French Open (1999)[15] and the Australian Open (2003).[16] Agassi also won seventeen ATP Masters Series titles, won the 1990 ATP Tour World Championships and was part of a winning Davis Cup team in 1990 and 1992.[5] He is the only male singles player in history to have won all four Grand Slam tournaments, the Olympic gold medal and the ATP Tour World Championships: a distinction dubbed as a "Career Super Slam" by *Sports Illustrated*.[17]

After suffering from sciatica caused by two bulging discs in his back, a spondylolisthesis (vertebral displacement) and a bone spur that interfered with the nerve, Agassi retired from professional tennis on September 3, 2006, after losing in the third round of the US Open. He is the founder of the Andre Agassi Charitable Foundation,[18] which has raised over $60 million for at-risk children in Southern Nevada.[19] In 2001, the Foundation opened the Andre Agassi College Preparatory Academy in Las Vegas, a K–12 public charter school for at-risk children.[20]

Agassi resides in Las Vegas, Nevada with his wife, retired professional tennis player Steffi Graf, and their two children.

Contents [hide]

1970–1985: Early life

[ed

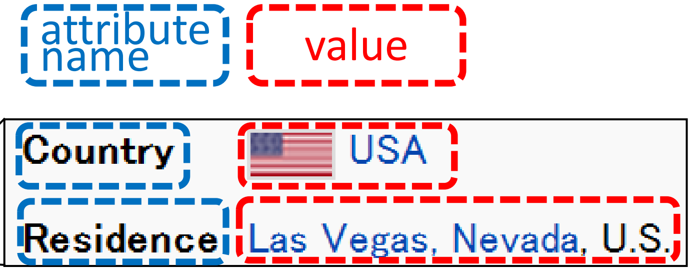
Agassi was born in Las Vegas, to Emmanuel "Mike" Aghassian and Elizabeth "Betty" Agassi (née Dudley).[3][21] His father was an Olympic boxer for Iran and of Assyrian and Armenian descent and when he migrated to the United States from Ira he changed his name to Agassi.[22][23][24][25][26] Andre Agassi's mother, Betty, is a breast cancer survivor. He has three older siblings – Rita, Philip and Tami.[27]

Mike Agassi reportedly banged on the fences with a hammer during Andre's matches when his son lost a point, screamed at officials and was ejected more than once.

In a passage from the book *Open*, Agassi details how his father made him play a match for money with football legend Jim Brown, in 1979, when Agassi was just 9 years old. Brown was at a Vegas tennis club complaining to the owner about a money match that was canceled. Agassi's father stepped in and told Brown that he could play his son and he would put up his house for the wager. Brown countered with a $10,000 bet; but after being warned by the club owner not to take the bet because he would lose and be embarrassed, Brown and Mike Agassi agreed that they'd set the amount after he and Andre played two sets. Brown lost those sets, 3–6, 3–6, declined the 10K wager, and offered to play the third for $500, which he lost 2–6.[28]

Country	USA
Residence	Las Vegas, Nevada, U.S.
Born	April 29, 1970 (age 41) Las Vegas, Nevada, U.S.
Height	1.80 m (5 ft 11 in)
Turned pro	1986
Retired	September 3, 2006
Plays	Right-handed (two-handed backhand)
Career prize money	US$31,152,975 • 4th All-time leader in earnings
Int. Tennis HOF	2011 (member page)
Singles	
Career record	870–274 (76.05%) (on the Grand Prix tour, ATP Tour, in Grand Slams and Davis Cup)
Career titles	60 according to the ATP and 68 in total
Highest ranking	No. 1 (April 10, 1995)
Grand Slam results	
Australian Open	W (1995, 2000, 2001, 2003)

http://en.wikipedia.org/wiki/Andre_Agassi

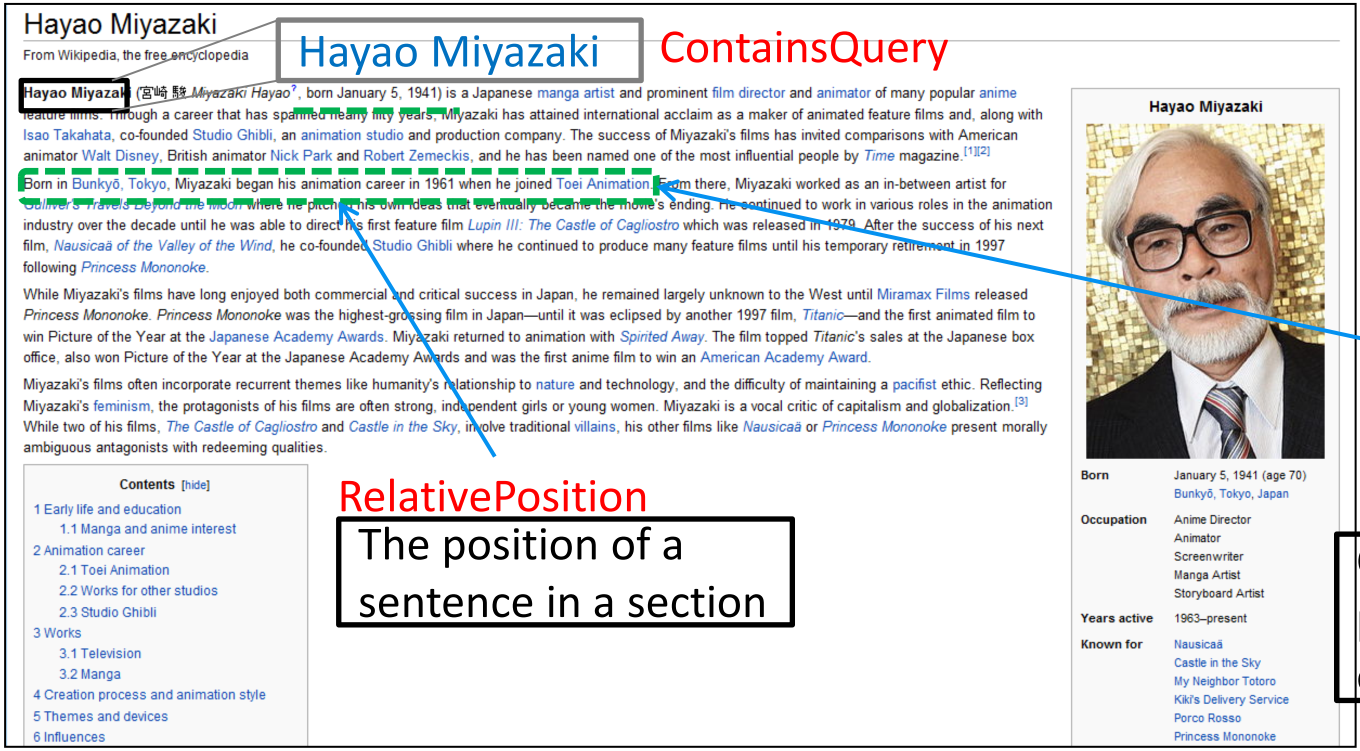


"Andre Agassi" ➡

➡ "Country is USA, residence is Las Vegas, Nevada, U.S., height is …., and highest ranking is No.1."

IE for CELEBLITY (Sentence-based)

- Extract important sentences from Wikipedia
- 12 features employed



Hayao Miyazaki

From Wikipedia, the free encyclopedia

Hayao Miyazaki (宮崎 駿 *Miyazaki Hayao*?, born January 5, 1941) is a Japanese manga artist and prominent film director and animator of many popular anime feature films. Through a career that has spanned nearly fifty years, Miyazaki has attained international acclaim as a maker of animated feature films and, along with Isao Takahata, co-founded Studio Ghibli, an animation studio and production company. The success of Miyazaki's films has invited comparisons with American animator Walt Disney, British animator Nick Park and Robert Zemeckis, and he has been named one of the most influential people by *Time* magazine.[1][2]

Born in Bunkyō, Tokyo, Miyazaki began his animation career in 1961 when he joined Toei Animation. From there, Miyazaki worked as an in-between artist for *Gulliver's Travels Beyond the Moon* where he pitched his own ideas that eventually became the movie's ending. He continued to work in various roles in the animation industry over the decade until he was able to direct his first feature film *Lupin III: The Castle of Cagliostro* which was released in 1979. After the success of his next film, *Nausicaä of the Valley of the Wind*, he co-founded Studio Ghibli where he continued to produce many feature films until his temporary retirement in 1997 following *Princess Mononoke*.

While Miyazaki's films have long enjoyed both commercial and critical success in Japan, he remained largely unknown to the West until Miramax Films released *Princess Mononoke*. *Princess Mononoke* was the highest-grossing film in Japan—until it was eclipsed by another 1997 film, *Titanic*—and the first animated film to win Picture of the Year at the Japanese Academy Awards. Miyazaki returned to animation with *Spirited Away*. The film topped *Titanic*'s sales at the Japanese box office, also won Picture of the Year at the Japanese Academy Awards and was the first anime film to win an American Academy Award.

Miyazaki's films often incorporate recurrent themes like humanity's relationship to nature and technology, and the difficulty of maintaining a pacifist ethic. Reflecting Miyazaki's feminism, the protagonists of his films are often strong, independent girls or young women. Miyazaki is a vocal critic of capitalism and globalization.[3] While two of his films, *The Castle of Cagliostro* and *Castle in the Sky*, involve traditional villains, his other films like *Nausicaä* or *Princess Mononoke* present morally ambiguous antagonists with redeeming qualities.

Contents [hide]

Hayao Miyazaki

Born	January 5, 1941 (age 70) Bunkyō, Tokyo, Japan
Occupation	Anime Director Animator Screenwriter Manga Artist Storyboard Artist
Years active	1963–present
Known for	Nausicaä Castle in the Sky My Neighbor Totoro Kiki's Delivery Service Porco Rosso Princess Mononoke

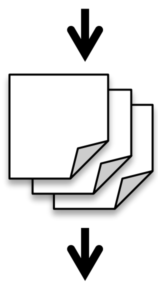
Hayao Miyazaki

ContainsQuery

RelativePosition

The position of a sentence in a section

query: Hayao Miyazaki



(Yahoo! Japan Web search API)

Summaries(Top n)

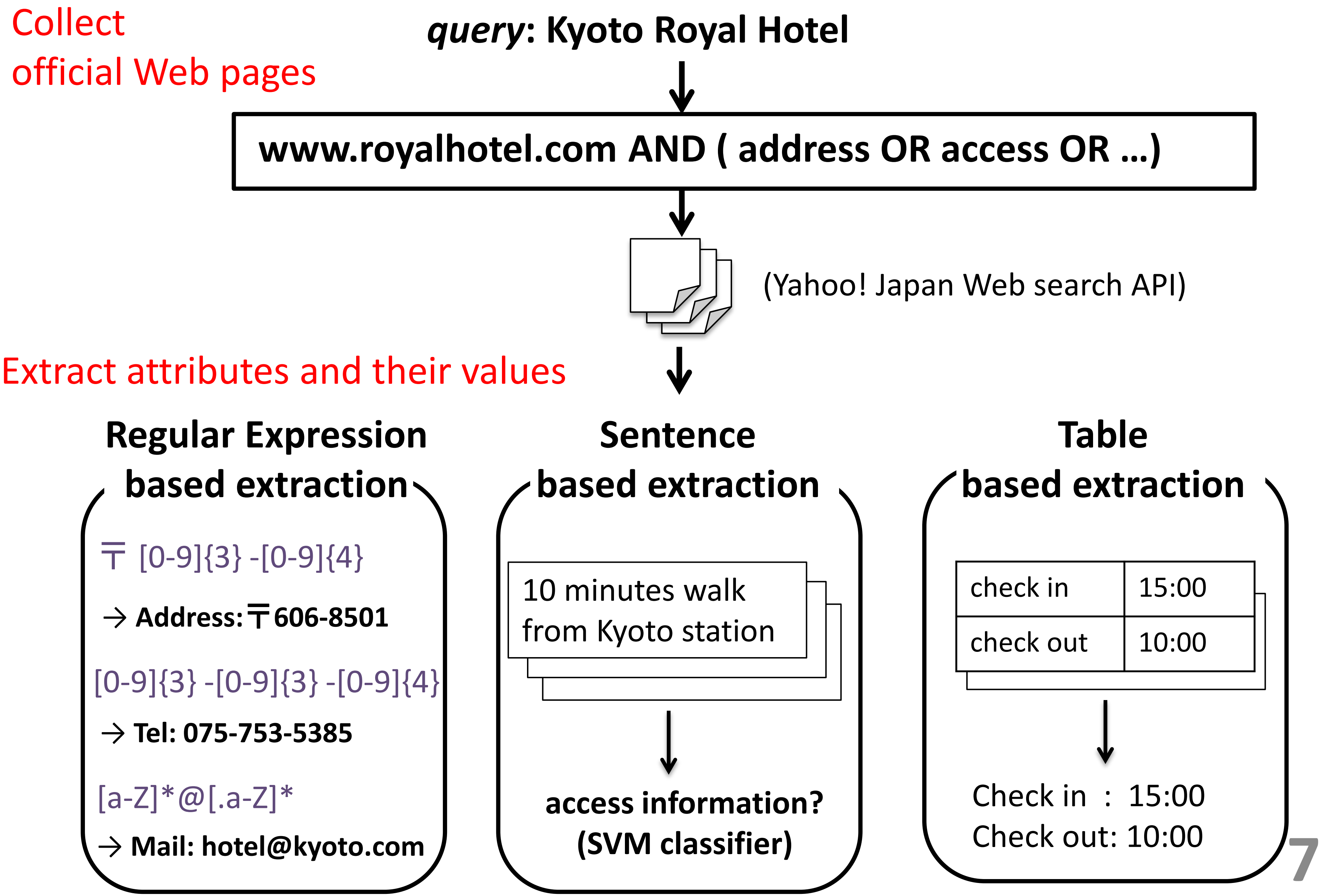
MaxCos, MinCos, MaxInnerProd, AvgCos, AvgInnerProd

Calculate cosine, inner-product similarity with each summary

http://en.wikipedia.org/wiki/Hayao_Miyazaki

- Train a regressor to predict the importance score of each sentence (using pre-distributed queries)

IE for LOCATION



Collect
official Web pages

query: Kyoto Royal Hotel

www.royalhotel.com AND (address OR access OR …)

(Yahoo! Japan Web search API)

Extract attributes and their values

Regular Expression based extraction

〒 [0-9]{3} -[0-9]{4}

→ **Address:〒606-8501**

[0-9]{3} -[0-9]{3} -[0-9]{4}

→ **Tel: 075-753-5385**

[a-Z]*@[.a-Z]*

→ **Mail: hotel@kyoto.com**

Sentence based extraction

10 minutes walk
from Kyoto station

access information?
(SVM classifier)

Table based extraction

check in	15:00
check out	10:00

Check in : 15:00
Check out: 10:00

IE for DEFINITION

- Query : *x*
- Search with a new query "*x towa*（とは）"
- Extract a sentence including the phrase of "*x towa*" from each retrieved Web page
- Apply the *LexRank*[1] algorithm to the set of sentences for estimating the sentence importance

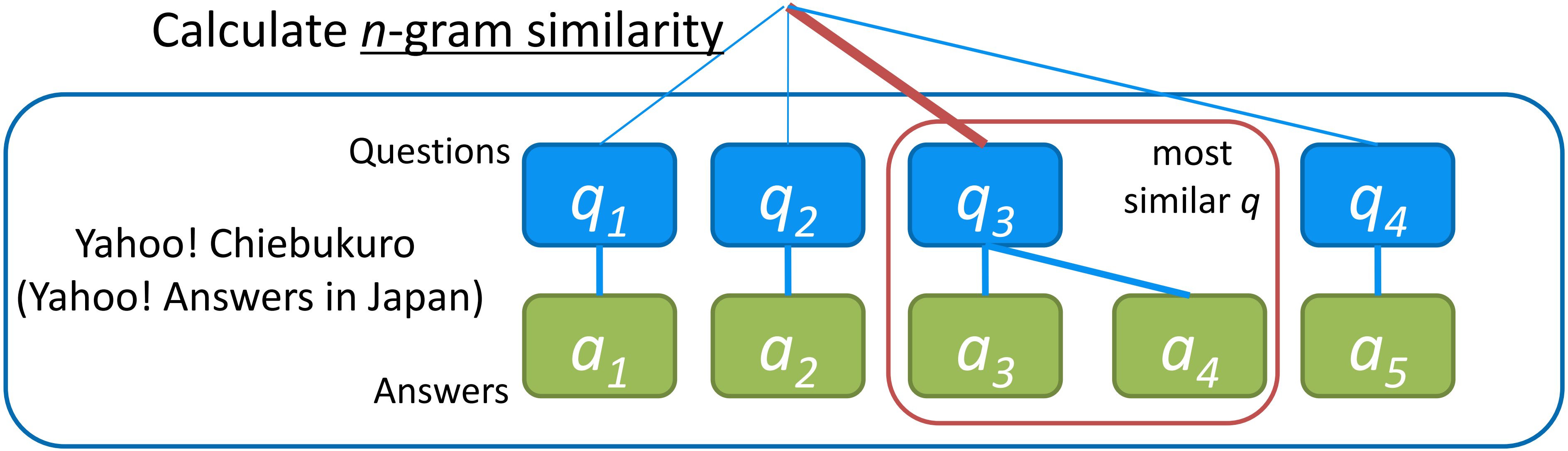
$$\mathbf{p} = \left[d\mathbf{U} + (1-d)\mathbf{B} \right]^{\mathrm{T}} \mathbf{p}$$

p : a vector representing the importance degree of each sentence
U : *n* × *n* square matrix whose elements are 1/*n*
B : adjacency matrix of the cosine similarity between two sentences
d : damping factor

[1]G. Erkan and R.D. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1), pages 457-479, 2004.

IE for QA

- Query: Which is taller, Tsutenkaku or Ustunomiya tower?



Calculate *n*-gram similarity

Questions

Yahoo! Chiebukuro
(Yahoo! Answers in Japan)

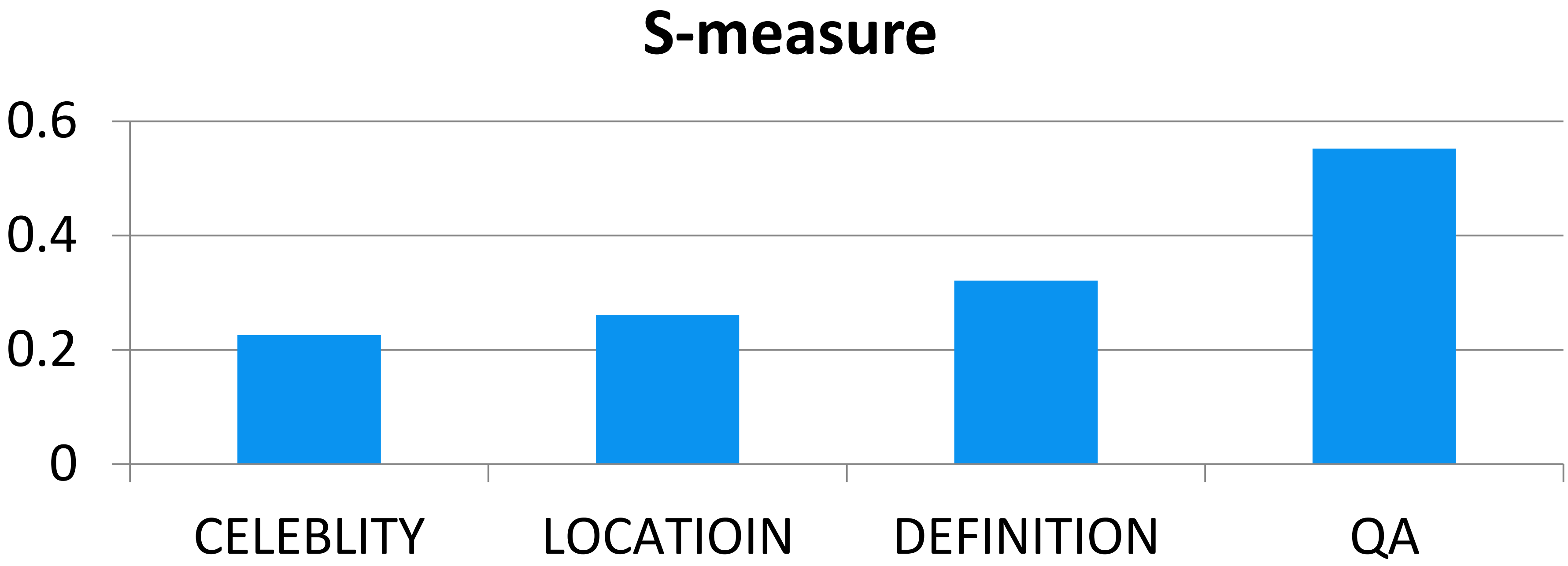
Answers

most similar *q*

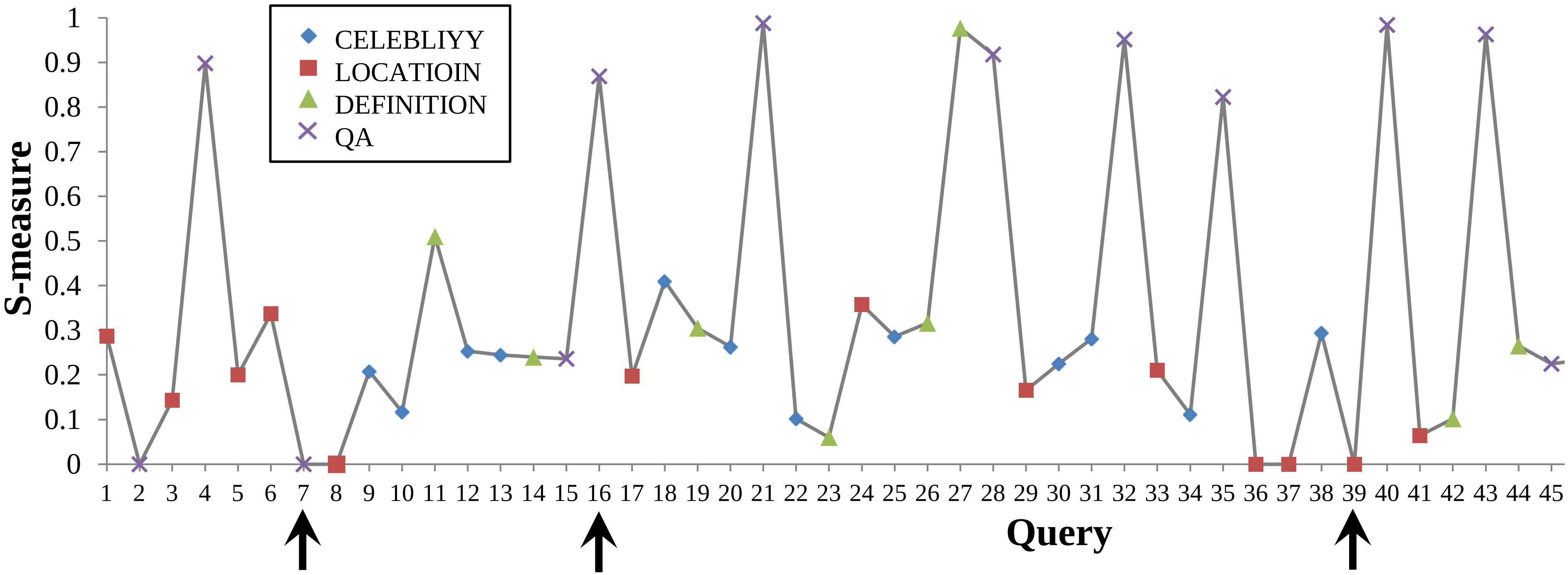
- Extract answers for most similar questions to the given query from Yahoo! Chiebukuro
- Return the best answer
  - if no best answer is given, return the most similar answer to the query

Results

- The accuracy of our query classifier: **0.93**
- **Highest S-measures for CELEBLITY and QA queries among participants**
  - – though the difference is not significant



Result for Each Query



Where is Yumura-onsen?

Which is taller, Tsutenkaku or Utsunomiya tower?

Kanazawa University

- For the 1CLICK task, the object identification problem should be tackled (see Kanazawa Univ.)
- QA queries obtained high S-measures when exactly the same questions are available (see the two questions)

Conclusions

- Information extraction framework for 1CLICK
  - Query classifier (**multiclass SVM**)
  - Four types of information extractor
    - CELEBRITY (**attribute and sentence extraction from Wikipedia**)
    - LOCATION (**postal and access information extraction**)
    - DEFINITION (**summarization of Web search results**)
    - QA (**QA pair extraction from Yahoo! Chiebukuro**)
  - Information summarizer (**MMR**)

Thanks! ntcir-9@dl.kuis.kyoto-u.ac.jp