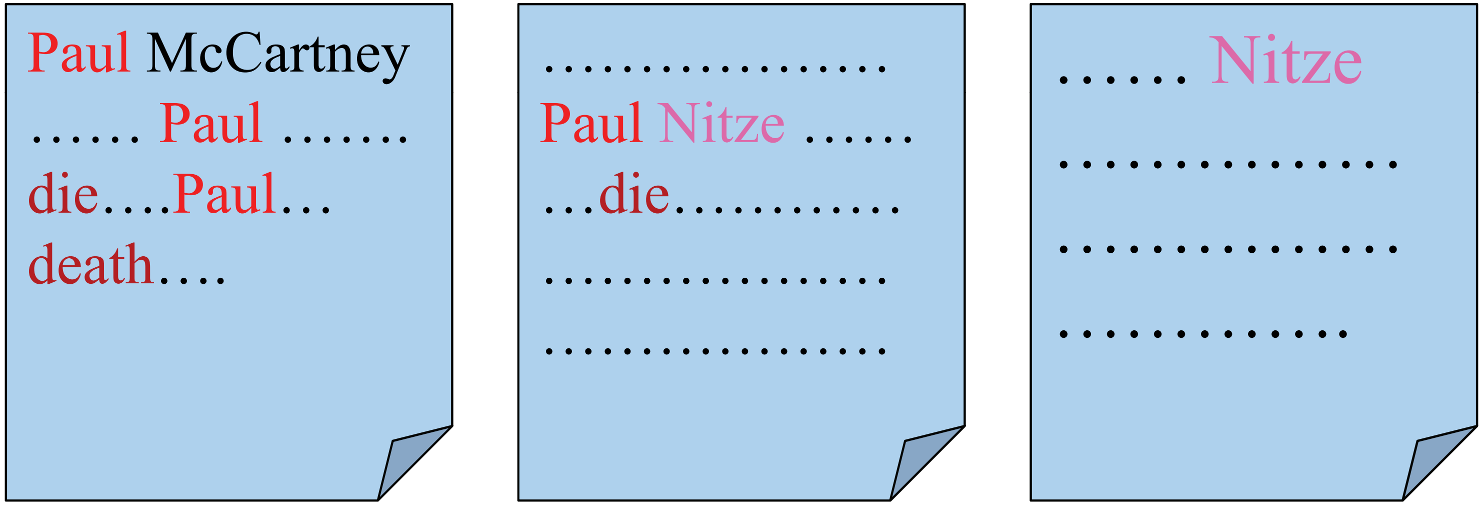
On a Combination of Probabilistic and Boolean IR Models for GeoTime Task

Masaharu YOSHIOKA
Hokkaido University

Motivation

- Information retrieval for Question Answering about a particular named entity
  - Documents that do not contain information about the named entity are irrelevant.

Query: When Paul Nitze die?



Paul McCartney
…… Paul …….
die….Paul…
death….

………………
Paul Nitze ……
…die…………
………………
………………

…… Nitze
……………
……………
…………

Score of IR System with Partial match	○	○	×
IR for QA	×	○	×

Proposed System

- Combination of Probabilistic and Boolean IR Models for QA
  - ABRIR (Appropriate Boolean query Reformulation for Information Retrieval)
    - Basic score is calculated by probabilistic IR model.
    - Documents that do not satisfy given Boolean query get penalty score.
  - Construction of appropriate Boolean Query for GeoTime QA
    - Verbs: synonym
    - Named entity: variation of description

ABRIR (a Probabilistic IR Model)

- Modified version of OKAPI
  - Use BM25 formula to calculate each document score

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

$$K = \frac{document\ length}{average\ document\ length}$$

$$w^{(1)} = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}$$

tf : frequency of T in a document
qtf : frequency of T in a query
k_1, k_3: parameter（k_1=1,k_3=1000 (initial) or 7 (final))
N: :the count of all documents in the database,
n: the count of all documents containing T
R: the given number of relevant documents
r : the count of all relevant documents containing T

  - Term weighting for phrasal terms
    - Document score may differ according to the dictionary entry

情報処理→ Word 情報処理
情報科学→ Word 情報, 科学 Phrase !c情報科学

    - Discount score for phrasal terms

$$qtf = c * qtf_c$$

qtf_c : frequency of phrase T in a query
c : parameter (c≤1;c=0.3)

Relevance Feedback

- Relevance feedback
  - Pseudo-relevance feedback
  - Query expansion
    - Use terms in relevant documents as query terms
    - Rocchio-type feedback

$$qtf = \alpha * qtf_0 + (1 - \alpha) * \frac{\sum_{i=1}^{R} qtf_i}{R}$$

qtf_0 : frequency of T in a initial query
qtf_i : frequency of T in a i-th relevant documents
R: the given number of relevant documents
α : parameter (α=0.7)

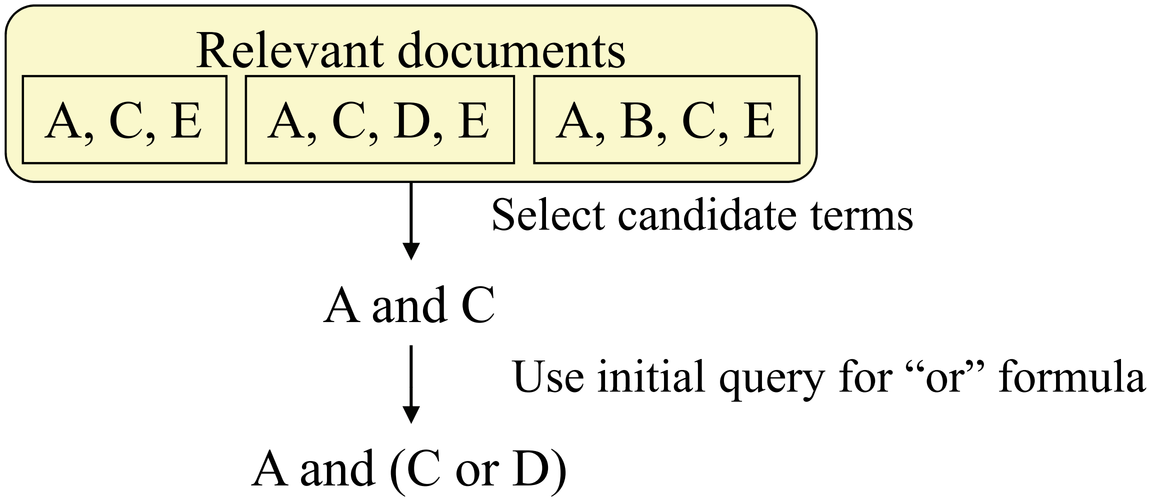
Problem on a Boolean IR model

- Retrieval performance of a Boolean IR model is worse than a probabilistic one
  - A Boolean query formula is expressive but is very difficult to construct appropriate one.
- Requirement for a Boolean query construction support
  - Use relevant documents for clarifying a Boolean query formula
    - Initial document retrieval without using a Boolean IR model
    - Relax a Boolean query formula by using relevant documents

Reconstruction of a Boolean Query Formula

- Relax an initial Boolean query formula to include given relevant documents as relevant one
  - Use terms that exists in all relevant documents and also exists in an initial query as a candidate to construct a relaxed Boolean query formula
  - Use an initial query for “or” formula

Initial query: (A and B and (C or D))



Relevant documents

A, C, E

A, C, D, E

A, B, C, E

Select candidate terms

A and C

Use initial query for “or” formula

A and (C or D)

Combination of Probabilistic and Boolean IR Models

- Two approach
  - Use a Boolean IR model first and calculate score of each retrieved document by using a probabilistic model
  - Use a probabilistic IR model first and apply penalty for documents that do not satisfy a Boolean query formula
    - Penalty is calculated by using term importance in BM25 $\beta \times w^{(1)} \times \frac{(k_3 + 1)qtf}{k_3 + qtf}$ β : parameter
    - Penalty is calculated for each “and” element
    - For “or” formula, use penalty of a term that has highest one among them.

Modification of ABRIR for QA

- Usage of verb as index terms
  - QA query may include verbs that are necessary to find relevant documents.
  - Relevant documents may include synonyms instead of original verbs.
- Handling named entities
  - Identification of named entity is important for constructing a Boolean query.
- Number of relevant documents
  - There may be only a few relevant documents for a query.
- Number of query expansion terms
  - Large number of query expansion terms may cause concept drift for QA

Synonym Set Construction for Verbs

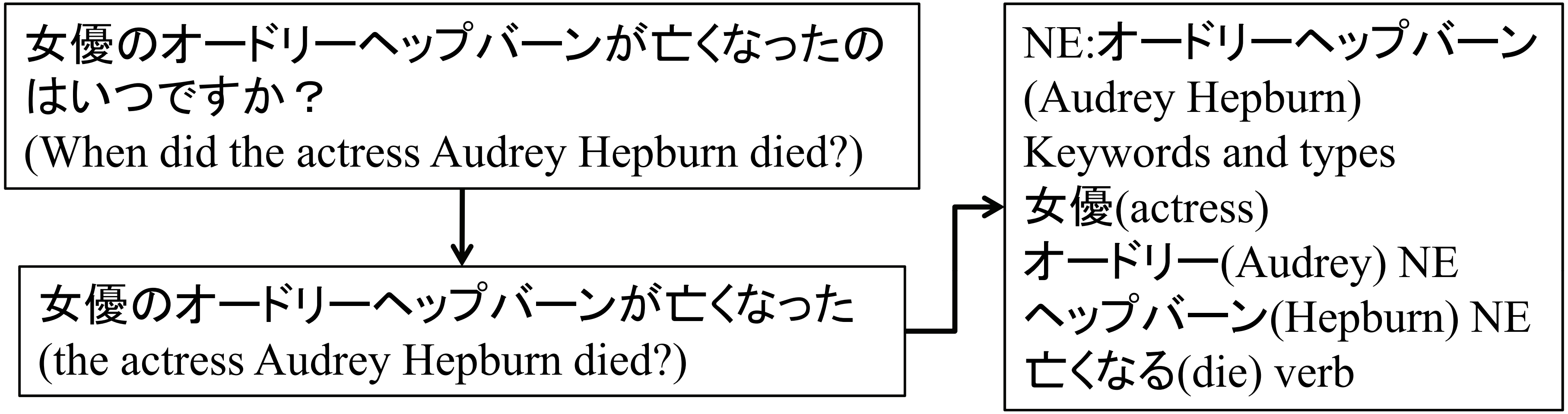
- Usage of Thesaurus (EDR) for synonym candidate generation.
- Usage of synonyms that exist in relevant documents are used for query expansion and a Boolean query formulation

Identification of Named Entity

- CaboCha is used for named entity extraction for query analysis
- Characteristic of named entity description in Japanese
  - Named entity of foreign people and organization may represent by using Katakana with some variation.
  - Construction of variation candidates by using simple candidate generation rule.
  - Usage of variations that exist in relevant documents are used for query expansion and a Boolean query formulation

Query Analysis and Boolean Query Reformulation

1. Remove question part of the query
   - Question part of the query (e.g., “のはいつですか？”(when)) is trimmed from the original query.
2. Morphological analysis and NE tagging
   - Almost same index terms extraction system is used for extract initial keywords.
     - Extraction of verb
     - Identification of named entities



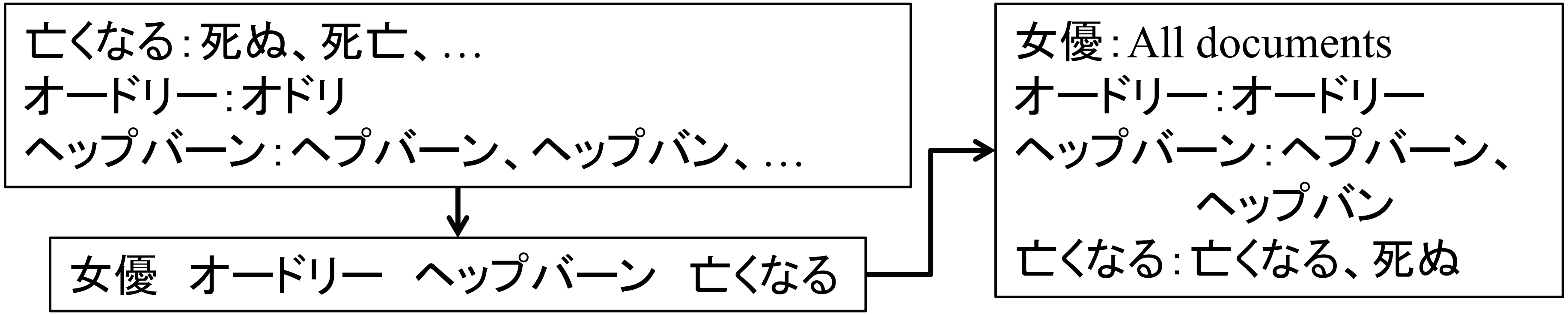
女優のオードリーヘップバーンが亡くなったのはいつですか？
(When did the actress Audrey Hepburn died?)

女優のオードリーヘップバーンが亡くなった
(the actress Audrey Hepburn died?)

NE:オードリーヘップバーン
(Audrey Hepburn)
Keywords and types
女優(actress)
オードリー(Audrey) NE
ヘップバーン(Hepburn) NE
亡くなる(die) verb

Query Analysis and Boolean Query Reformulation

3. Generation of synonym and variation list
   - The system generates synonym list for verbs and variation list for named entity.
4. Initial retrieval
   - Probabilistic IR model is used for finding out pseudo relevant documents. Top 3 ranked documents as pseudo relevant ones.
5. Comparison of pseudo relevant documents and query terms



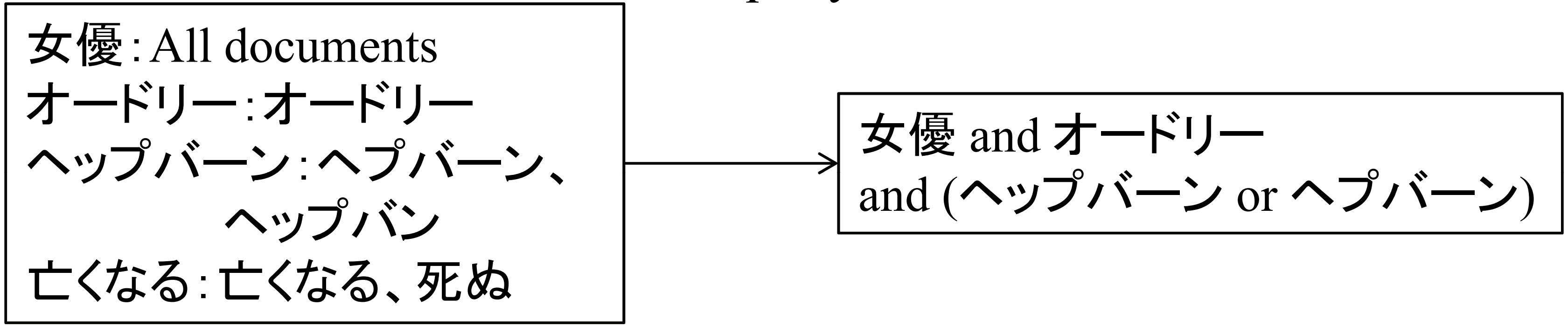
亡くなる:死ぬ、死亡、…
オードリー:オドリ
ヘップバーン:ヘプバーン、ヘップバン、…

女優　オードリー　ヘップバーン　亡くなる

女優:All documents
オードリー:オードリー
ヘップバーン:ヘプバーン、ヘップバン
亡くなる:亡くなる、死ぬ

Query Analysis and Boolean Query Reformulation

6. Construction of Boolean query
   - There are three types of keywords in query; e.g., NE, verb, and other. The system compares query keywords and pseudo relevant documents in following manner.
     - Named entity
       - Candidates that exist in the pseudo relevant documents are connected with “or”.
     - Other keywords in initial query
       - When other keywords in initial query exist in all pseudo relevant documents, These keywords are used as AND elements of the final query.

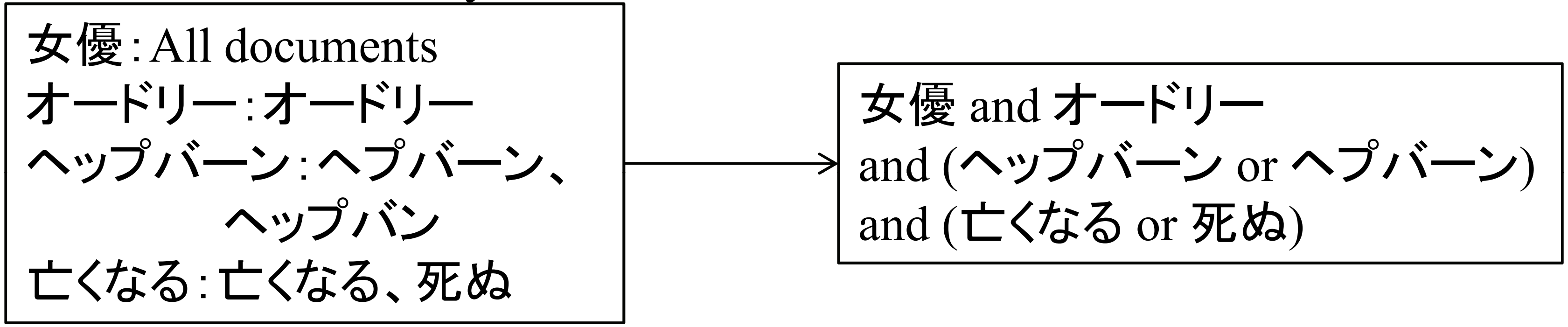


女優 : All documents
オードリー : オードリー
ヘップバーン : ヘプバーン、
ヘップバン
亡くなる : 亡くなる、死ぬ

女優 and オードリー
and (ヘップバーン or ヘプバーン)

Query Analysis and Boolean Query Reformulation

6. Construction of Boolean query
   - Verb
     - When all pseudo relevant documents contains one or more synonyms of the verb, these documents are sufficient enough for generating synonym list for final Boolean query.
     - Secondary retrieval (option)
       - When there is one or more document(s) that do not contain any synonyms, the system generates new query by replacing the verb with synonym list and conducts secondary retrieval.



女優: All documents
オードリー: オードリー
ヘップバーン: ヘプバーン、
ヘップバン
亡くなる: 亡くなる、死ぬ

女優 and オードリー
and (ヘップバーン or ヘプバーン)
and (亡くなる or 死ぬ)

Query Analysis and Boolean Query Reformulation

7. Query expansion by using pseudo relevant documents
   - The system selected the 5 different terms with the highest mutual information content between a relevant document set and a term. The system also add keywords in Boolean query as expansion terms.
8. Final retrieval
   - Final retrieval is conducted by the probabilistic IR model. The Boolean query is used for filtering out the document or calculating penalty for the document

Experimental Result

- Parameters for the probabilistic IR model
  - $k_1 = 1$, $k_3 = 7$, $K = dl/avdl$, $c = 0.3$, $\alpha = 0.7$ *dl is the length of a document* (the number of terms and phrasal terms) *avdl is the* average length of all documents.
- Parameter for penalty calculation
  - $\beta = 3$
  - $\beta_n = 1000000$ (penalty for named entity)

Submitted Runs

Runs	Type	Boolean for NE	Boolean for others	
HU-KB-JA-JA-01-D	Description	Filter	Filter	
HU-KB-JA-JA-02-DN	Description + Narrative	Filter	Filter	
HU-KB-JA-JA-03-D	Description	Penalty	Penalty	
HU-KB-JA-JA-04-D	Description	Penalty	No	
HU-KB-JA-JA-05-D	Description	No	No	Baseline Okapi

HU-KB-JA-JA-03-D
ABRIR with Boolean penalty is the best system with Description

Evaluation Measure

- Average over all topics
  - 03-D is better than 01-D
    - Boolean penalty is better than Boolean filter
  - All runs with Boolean query outperforms baseline Okapi

run	01-D	02-DN	03-D	04-D	05-D
AP	0.3697	0.3867	0.3719	0.3627	*0.2881*
nDCG	0.4117	0.4268	0.4162	0.4072	*0.3282*
Q	0.5710	0.5685	0.5881	0.5717	*0.4993*

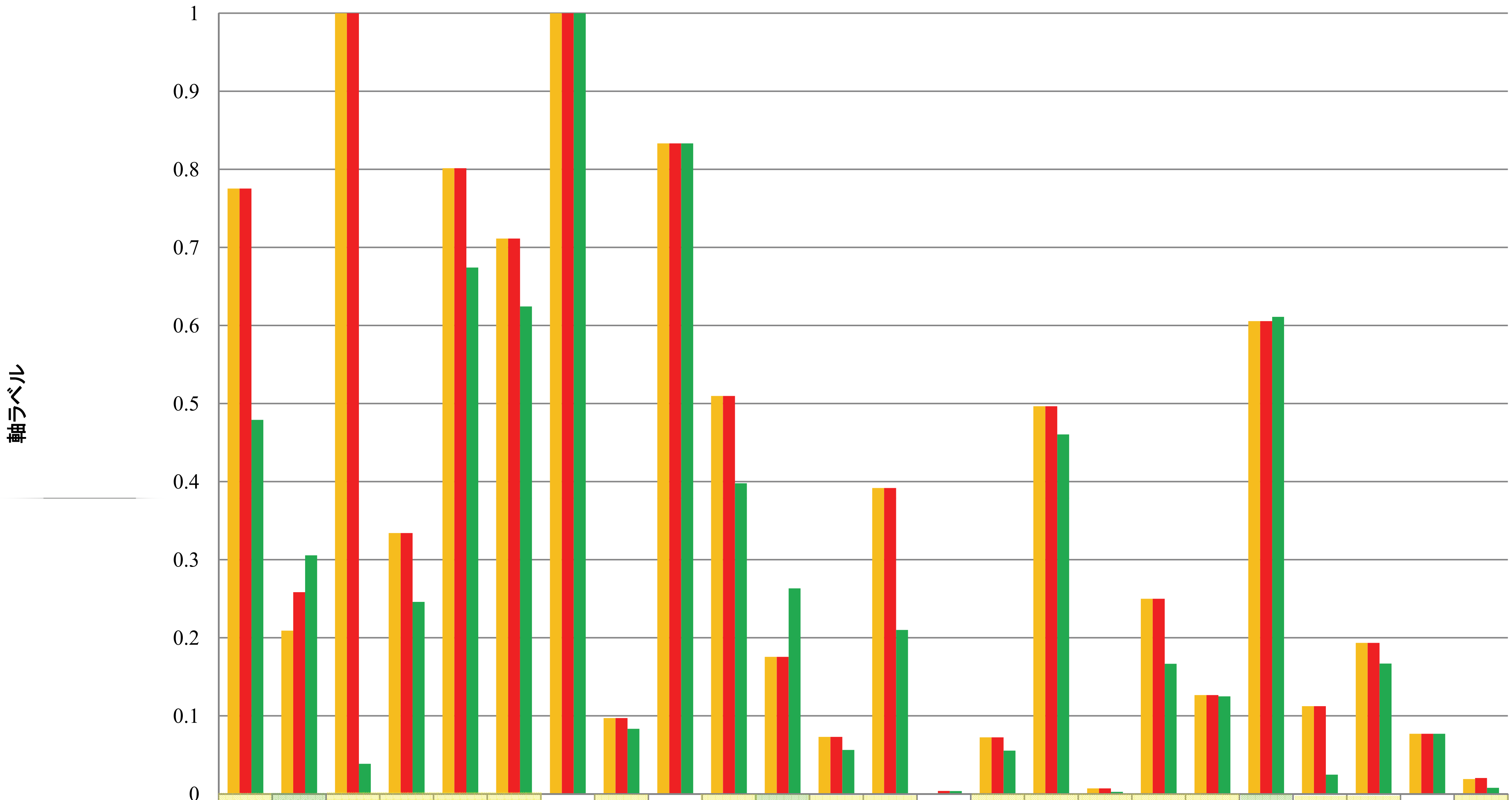
Appropriateness of Boolean Query for Filter

- Boolean filter to restrict candidate documents less than 1000 (10 topics).
  - Appropriate Boolean query (all relevant documents match the constructed Boolean query): 7 topics (Topic:Boolean matched documents; 1:772, 3:1, 6:275, 9:6,13:105, 19:329, 20:945)
  - Inappropriate Boolean query (some relevant documents do not match the query): 3 topics (Topic:filter out/total rel 2:26/48, 14: 2/2, 25: 1/3)

Statistical Significance Test between 03-D and 05-D (Baseline: Okapi)

- t-Test with significance level of 0.05 for a two-sided tests
    - Statistically significant
        - nDCG(0.018) and Q(0.040)
    - Not Statistically significant
        - AP(0.055)
- Wilcoxon Signed Rank test
    - Statistically significant
        - AP(0.0015), nDCG(0.0006) and Q(0.0024)

AP Topic by Topic



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	20	21	22	23	24	25
🟧 HU-KB-JA-JA-01-D	0.775	0.209	1	0.334	0.801	0.711	1	0.097	0.833	0.51	0.176	0.073	0.392	0	0.072	0.497	0.007	0.25	0.127	0.606	0.112	0.193	0.077	0.019
🟥 HU-KB-JA-JA-03-D	0.775	0.258	1	0.334	0.801	0.711	1	0.097	0.833	0.51	0.176	0.073	0.392	0.004	0.072	0.497	0.007	0.25	0.127	0.606	0.112	0.193	0.077	0.02
🟩 HU-KB-JA-JA-05-D	0.479	0.306	0.039	0.246	0.674	0.624	1	0.083	0.833	0.398	0.263	0.056	0.21	0.004	0.055	0.461	0.003	0.167	0.125	0.611	0.025	0.167	0.077	0.008

Failure Analysis

- Topic2
  - “ハリケーン”(hurricane) is recognized as named entity
  - “ハリケーン” (hurricane) without “カトリーナ” (Katrina) get similar score “カトリーナ” (Katrina) without “ハリケーン” (hurricane).
- Topic 11 and 21
  - Those topics don’t contain the named entity information. Topic 11 is a survey type question and is different from assumption of the question.
- Topic 14
  - Named entity keyword “アフリカ” (Africa).
  - However, the relevant documents has name of the African country “コンゴ民主共和国” (Democratic Republic of the Congo) instead of “アフリカ”.

Conclusion

- Proposal of using ABRIR as an IR system for question and answering for particular named entities.
- From the evaluation experiment, we confirm that ABRIR can make appropriate Boolean query and penalty based system outperform the baseline system (probabilistic IR model: Okapi BM25).

Future Works

- Consideration of different method to generate variation list for named entity.
  - Usage of Wikipedia redirect
  - Application of transliteration method
  - Part of relationship (Africa and Congo)
- Application of this approach to Web documents
  - There are more varieties of description for named entity in Web documents.