Overview of the NTCIR-10 SpokenDoc-2 Task

Tomoyosi Akiba (Toyohashi University of Technology)
Hiromitsu Nishizaki (Yamanashi University)
Kiyoaki Aikawa (Tokyo University of Technology)
Xinhui Hu (National Institute of Information and Communications Technology)
Yoshiaki Itoh (Iwate Prefectural University)
Tatsuya Kawahara (Kyoto University)
Seiichi Nakagawa (Toyohashi University of Technology)
Hiroaki Nanjo (Ryukoku University)
Yoichi Yamashita (Ritsumeikan University)

What is “SpokenDoc-2”?

- Second round of the IR for spoken documents
- Finding the information related to a given query from speech data

Speech data ≅ Transcribed text by automatic speech recognition

“information retrieval”

Information retrieval (word-unit transcription)
In for ma tion re triev al (sub-word-unit)

Problem Including a lot of errors!!

Participants of SpokenDoc-2 will challenge

Information retrieval from very noisy text data

Techniques for SpokenDoc may be used for OCR or Machine Translated text retrieval

SpokenDoc-1 vs. SpokenDoc-2

- Previous evaluation frameworks related to Spoken Document Retrieval
  - TREC SDR Track (1996-2000), TREC Video Track (2001-2002), TRECVID (2003-2010)
  - CLEF CL-SDR (2003-2004), CLEF CL-SR (2005-2007), CLEF QAST (2007-2009), VideoCLEF (2008-2009), Mediaeval (2010-2011)
  - NIST STD evaluation (2006)
- NTCIR-9 SpokenDoc (2011)
  - Both STD and SCR task
  - First evaluation targeting Japanese and lecture speech
  - Investigate Boundary-free passage retrieval task
- NTCIR-10 SpokenDoc-2 (2013)
  - New target documents "Corpus of Spoken Document Processing Workshop (SDPWS)"
  - New "inexistent spoken term detection" (iSTD) task
  - Prepared Separate query topics for lecture and passage retrieval tasks.

Outline

- ✓ Background
- Task Definition
  - Documents & Transcriptions
  - Subtasks
- Evaluation Results
  - STD subtask
  - SDR subtask

Document Collection

- Corpus of Spontaneous Japanese (CSJ)
  - provided from National Institute for Japanese Language and Linguistics (NINJAL)
  - 2702 lectures, 612 hours length
- Recording of Annual Spoken Document Processing Workshop (SDPWS) *new*
  - 104 oral presentation, 28.6 hours length

Reference Automatic Transcriptions for CSJ and SDPWS

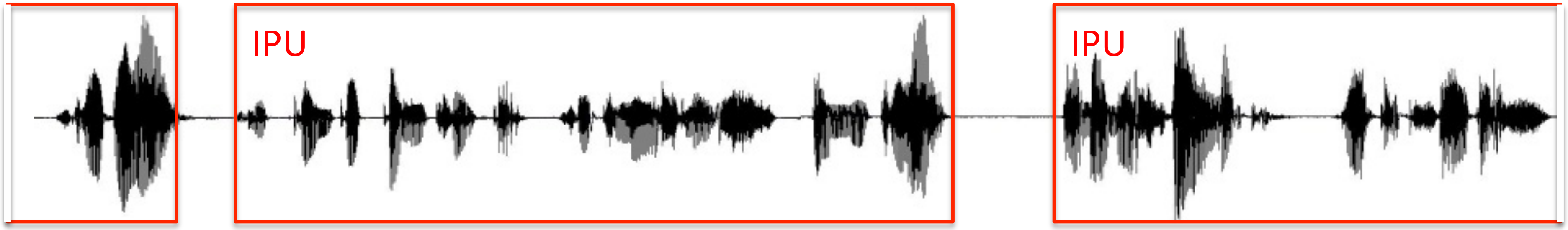
- Two Types of Recognition Unit
  - **Word-based transcription**
    - Produced by using the word tri-gram language model (Vocabulary Size is 27K)
  - **Syllable-based transcription**
    - Produced by using the syllable tri-gram language model (with Japanese Syllable Dictionary)
- Two Kinds of Training Condition
  - **Matched Condition**
    - Both the acoustic model and the language model were trained by using lecture speech (CSJ).
  - **Unmatched Condition**
    - The acoustic model was trained by lecture speech, while the language model was trained by newspaper articles.
- All of them are provided as N-best lists, confusion networks and word-lattices to represent the multiple recognition candidates

	word-based	**syllable-based**
matched condition	REF-WORD-MATCHED	REF-SYLLABLE-MATCHED
unmatched condition	REF-WORD-UNMATCHED	REF-SYLLABLE-UNMATCHED

The organizers provided 2 × 2 × 2 = 8 transcriptions in total.

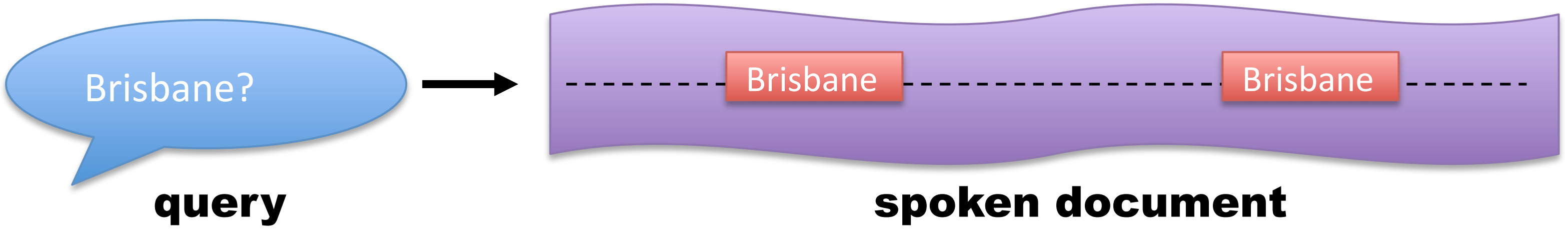
Inter Pausal Unit (IPU)

- Speech segment surrounded between two pauses no shorter than 200ms.
  - A spoken document = a sequence of IPUs
- Used as an atomic unit in our task definition.
  - Ignore time differences within IPU
- Enable us to apply conventional IR metrics based on discrete units.

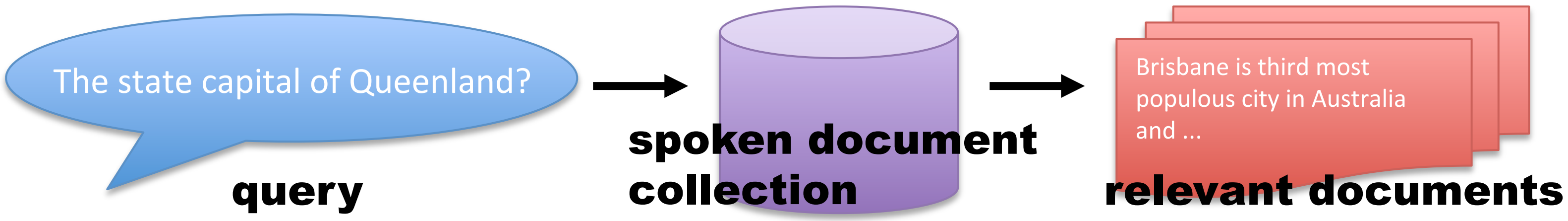


Task Definition Overview:

- Spoken Term Detection (STD) task
  - Find the occurrence of the given query term

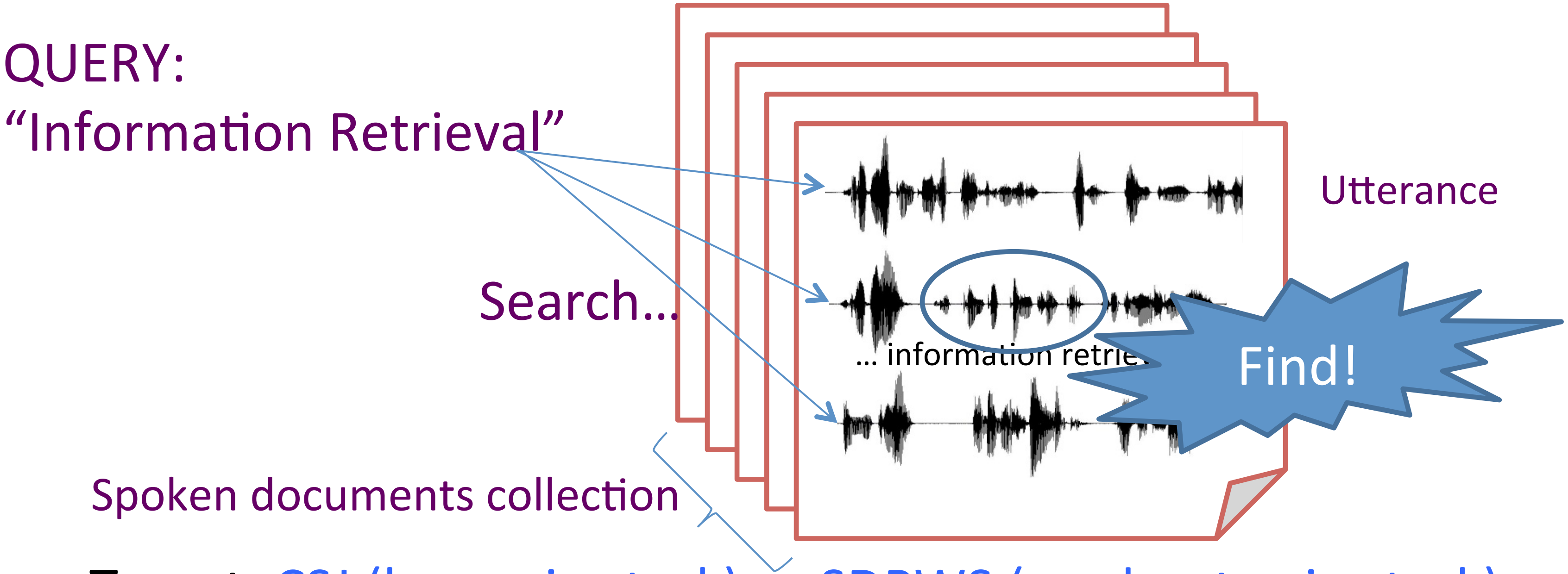


- Spoken Content Retrieval (SCR) task
  - Find the segments related to the given query topic


The state capital of Queenland?
query
spoken document collection
Brisbane is third most populous city in Australia and ...
relevant documents

Spoken term detection task (STD)

- From the target documents, the system is required to find the IPUs in that the given query term is uttered.



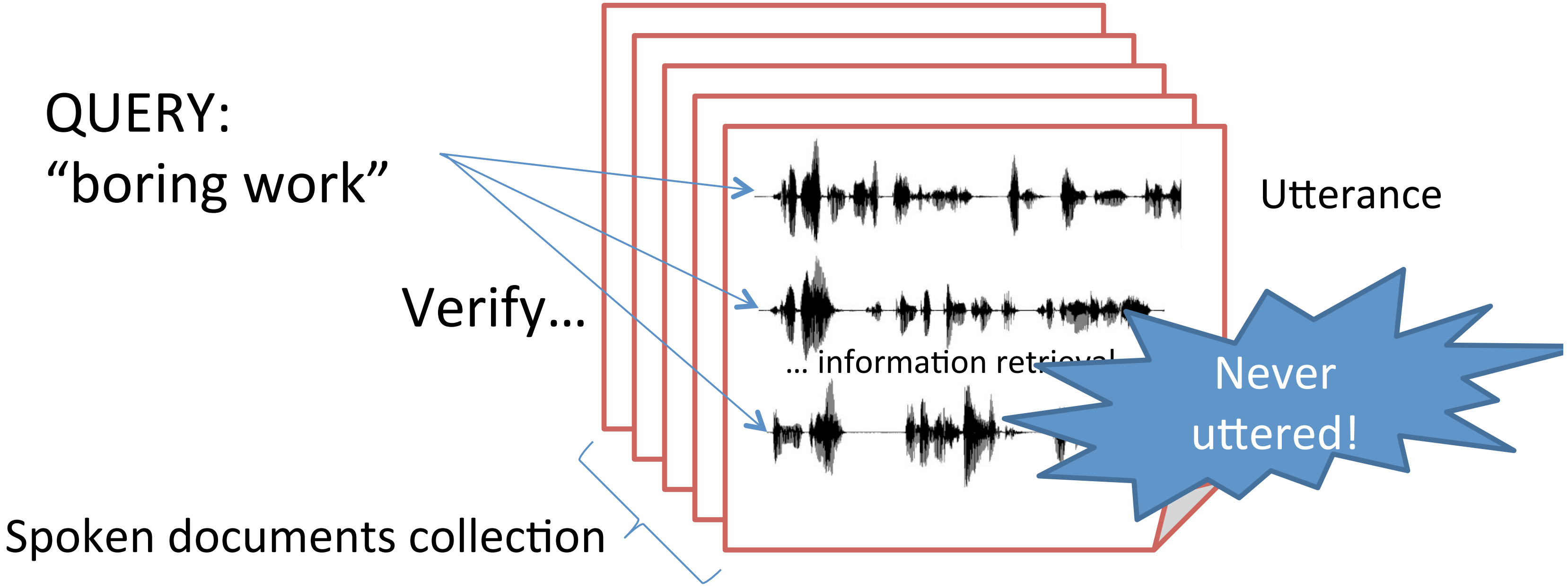
  - Target: CSJ (large-size task) or SDPWS (moderate-size task)
  - Input: a query term
  - Output: a list of scored Inter Pausal Units (IPUs)

Evaluation Metrics for STD task

- Using IPUs as the basic unit.
  - Precision, Recall, and F-measure (micro/macro average)
    - Recall-precision curve
    - F-measure at the specified detection threshold (actual F-measure)
    - max F-measure
  - Mean Average Precision (macro average)

Inexistent spoken term detection task (iSTD) *new*

- Which query terms are not uttered in a given spoken document collection?



QUERY:
“boring work”

Verify…

Utterance

... information retrieval

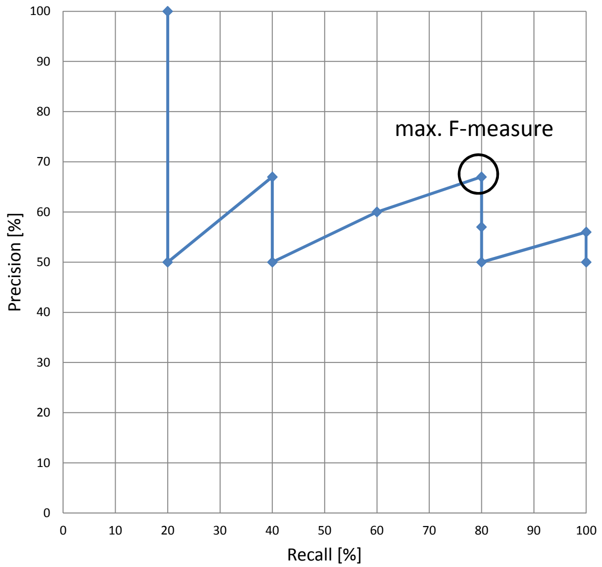
Never uttered!

Spoken documents collection

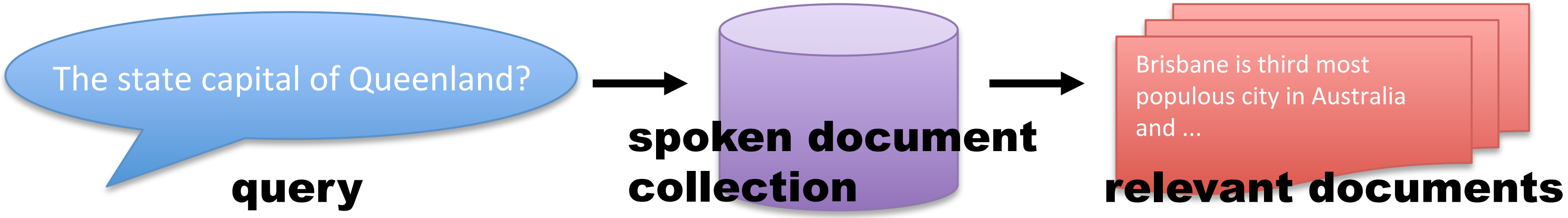
  - Target: SDPWS
  - Input: a list of query terms
  - Output: a ranked list of "not uttered" query terms

Evaluation Metrics of iSTD task

- Recall, Precision, and F-measure at Top-N of the ranked list.
  - F-measure at N=100.
  - F-measure at run specified N.
  - Maximum F-measure among N.
  - Recall-Precision curve.



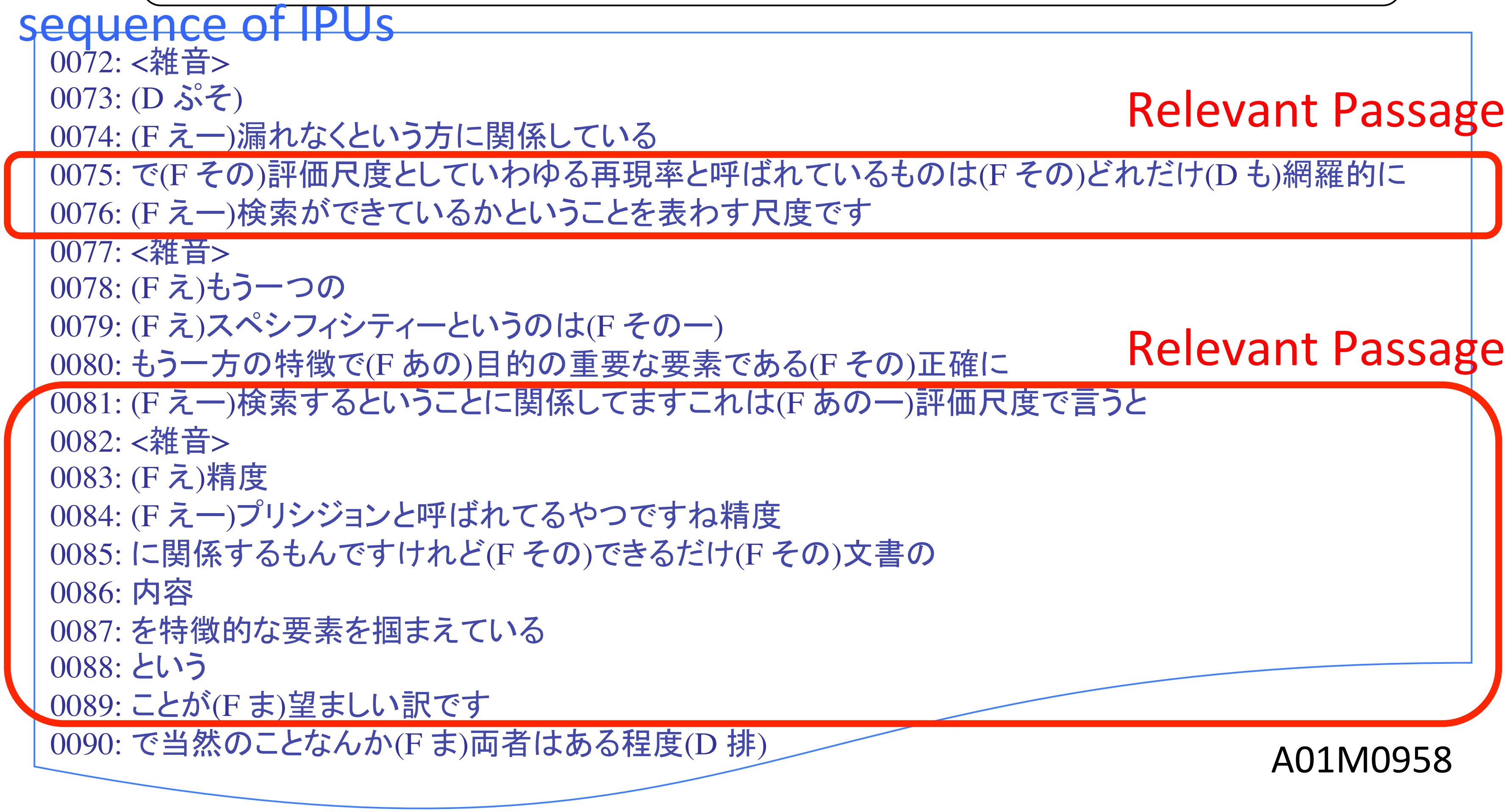
Task Definition of SCR task


The state capital of Queenland?
query
spoken document collection
Brisbane is third most populous city in Australia and ...
relevant documents

- From the target documents, the system is required to find the segments that are relevant to the given query topic.
  - Granularity of a segment: lecture or passage
  - Target: CSJ (lecture retrieval task) or SDPWS (passage retrieval task)
  - Input: a query topic
  - Output: an ordered list of segments

An Example Query Topic and its Relevant Passages

Query: 情報検索性能を評価するにはどのような方法があるか知りたい。
(How can we evaluate the performance of information retrieval?)



sequence of IPUs

0072: <雑音>
0073: (D ぷそ)
0074: (F えー)漏れなくという方に関係している
0075: で(F その)評価尺度としていわゆる再現率と呼ばれているものは(F その)どれだけ(D も)網羅的に
0076: (F えー)検索ができているかということを表わす尺度です
0077: <雑音>
0078: (F え)もう一つの
0079: (F え)スペシフィシティーというのは(F そのー)
0080: もう一方の特徴で(F あの)目的の重要な要素である(F その)正確に
0081: (F えー)検索するということに関係してますこれは(F あのー)評価尺度で言うと
0082: <雑音>
0083: (F え)精度
0084: (F えー)プリシジョンと呼ばれてるやつですね精度
0085: に関係するもんですけれど(F その)できるだけ(F その)文書の
0086: 内容
0087: を特徴的な要素を掴まえている
0088: という
0089: ことが(F ま)望ましい訳です
0090: で当然のことなんか(F ま)両者はある程度(D 排)

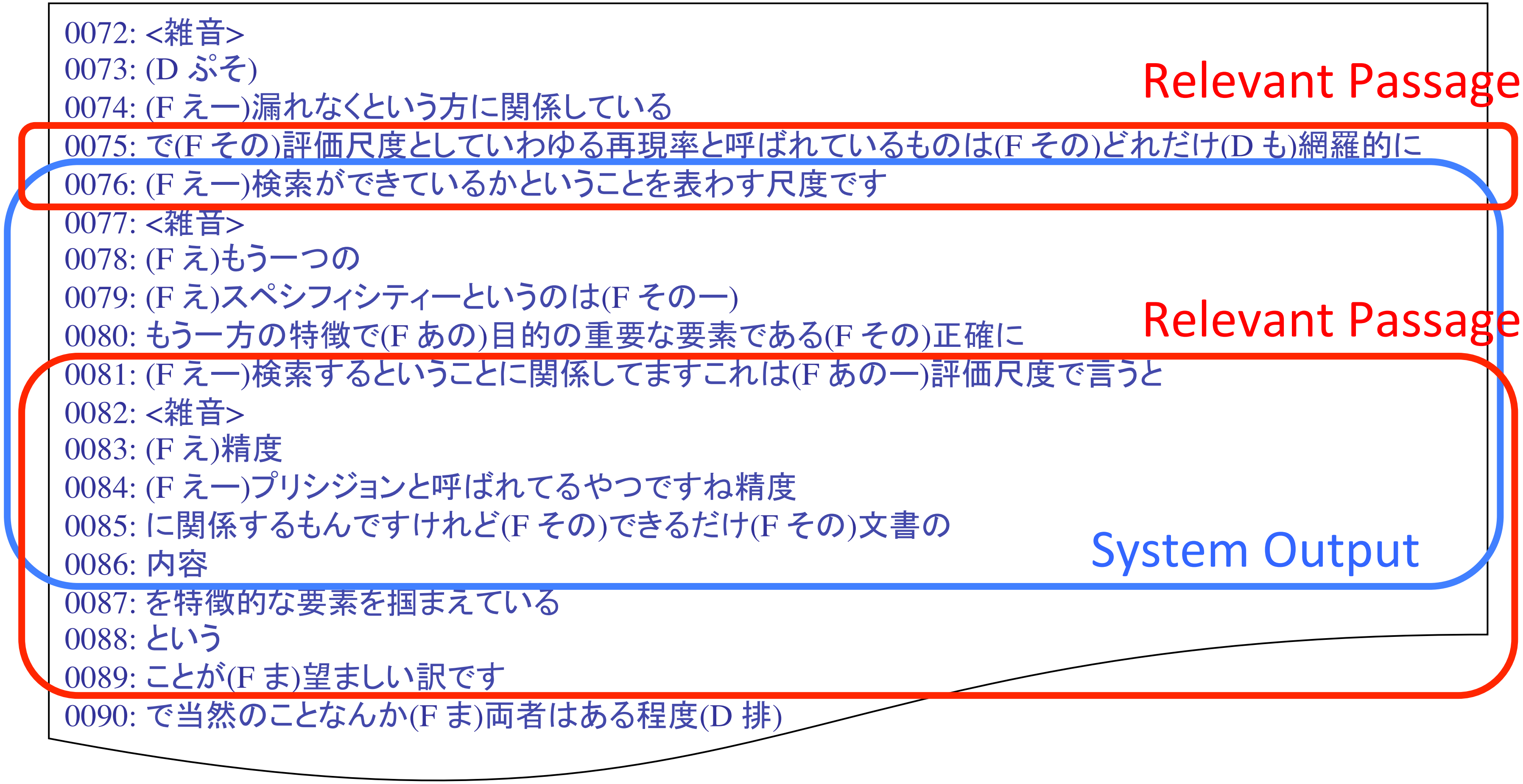
A01M0958

Evaluation Metrics for SCR task

- For lecture retrieval task
  - Mean Average Precision (MAP)
- For passage retrieval task
  - Utterance-based metric
    - utterance-based MAP (uMAP)
  - Passage-based metric
    - point-wise MAP (pwMAP)
    - fractional MAP (fMAP)

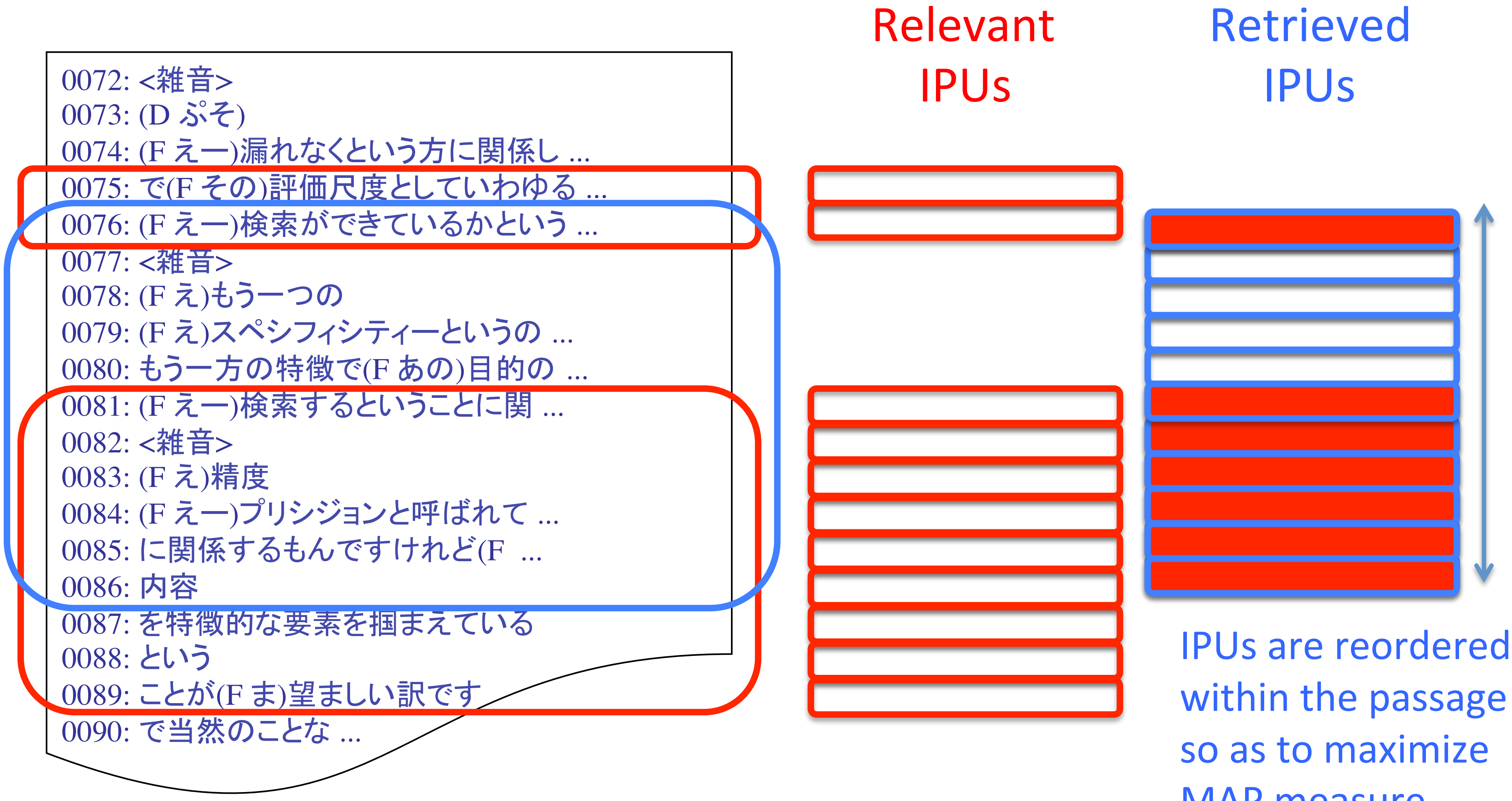
Evaluation of Passage Retrieval

Q:情報検索性能を評価するにはどのような方法があるか知りたい。



0072: <雑音>
0073: (D ぷそ)
0074: (F えー)漏れなくという方に関係している
0075: で(F その)評価尺度としていわゆる再現率と呼ばれているものは(F その)どれだけ(D も)網羅的に
0076: (F えー)検索ができているかということを表わす尺度です
0077: <雑音>
0078: (F え)もう一つの
0079: (F え)スペシフィシティーというのは(F その一)
0080: もう一方の特徴で(F あの)目的の重要な要素である(F その)正確に
0081: (F えー)検索するということに関係してますこれは(F あの一)評価尺度で言うと
0082: <雑音>
0083: (F え)精度
0084: (F えー)プリシジョンと呼ばれてるやつですね精度
0085: に関係するもんですけれど(F その)できるだけ(F その)文書の
0086: 内容
0087: を特徴的な要素を掴まえている
0088: という
0089: ことが(F ま)望ましい訳です
0090: で当然のことなんか(F ま)両者はある程度(D 排)

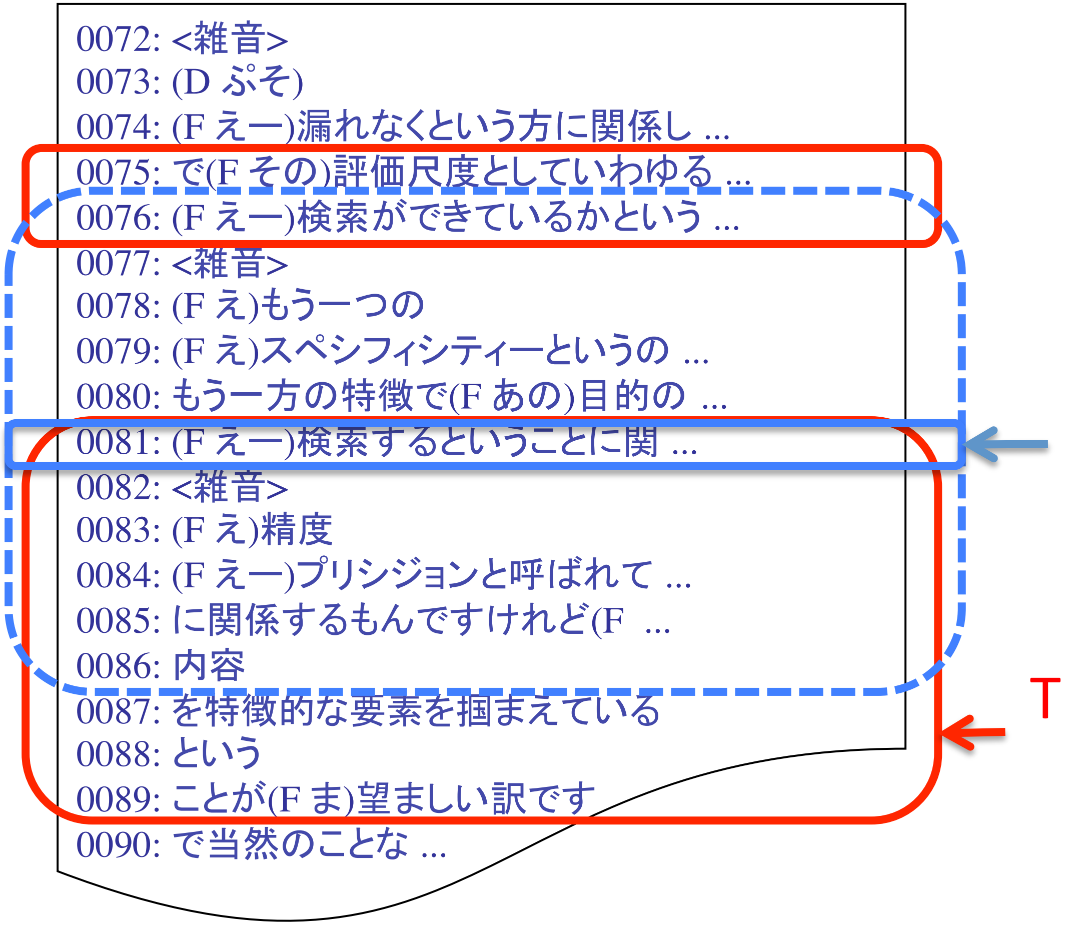
Utterance-based MAP (uMAP)



0072: <雑音>
0073: (D ぷそ)
0074: (F えー)漏れなくという方に関係し ...
0075: で(F その)評価尺度としていわゆる ...
0076: (F えー)検索ができているかという ...
0077: <雑音>
0078: (F え)もう一つの
0079: (F え)スペシフィシティーというの ...
0080: もう一方の特徴で(F あの)目的の ...
0081: (F えー)検索するということに関 ...
0082: <雑音>
0083: (F え)精度
0084: (F えー)プリシジョンと呼ばれて ...
0085: に関係するもんですけれど(F ...
0086: 内容
0087: を特徴的な要素を掴まえている
0088: という
0089: ことが(F ま)望ましい訳です
0090: で当然のことな ...

IPUs are reordered within the passage so as to maximize MAP measure.

Point-wise MAP (pwMAP)

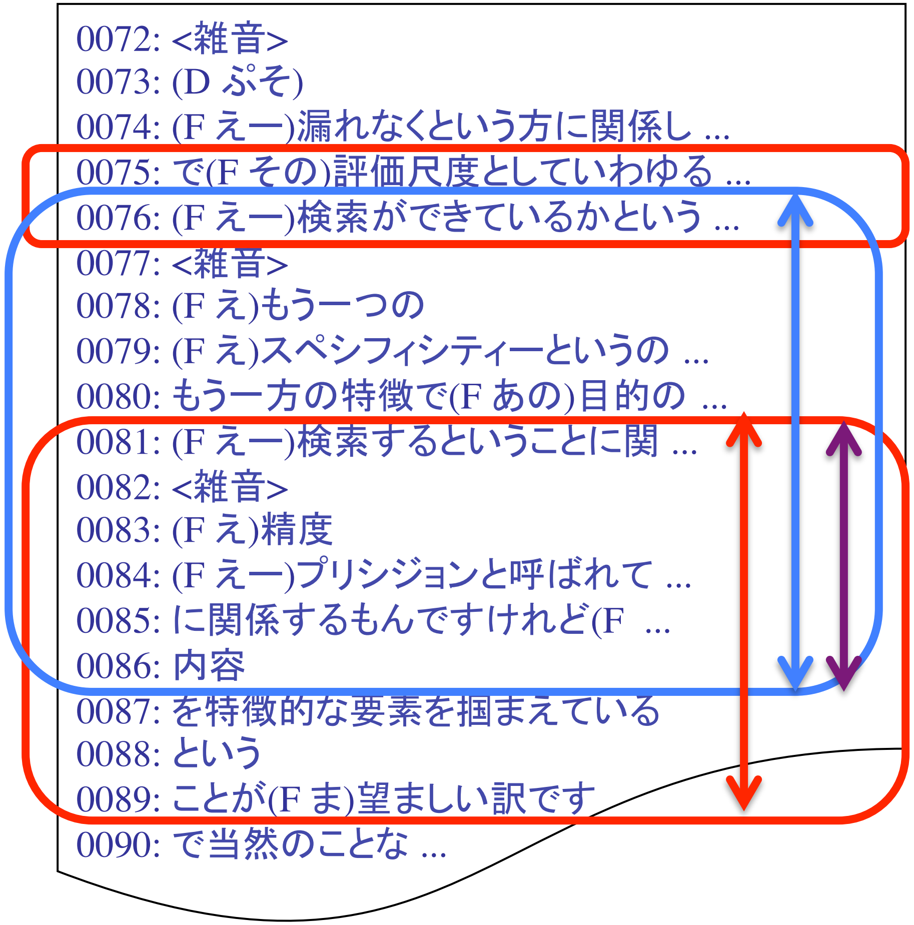


0072: <雑音>
0073: (D ぷそ)
0074: (F えー)漏れなくという方に関係し ...
0075: で(F その)評価尺度としていわゆる ...
0076: (F えー)検索ができているかという ...
0077: <雑音>
0078: (F え)もう一つの
0079: (F え)スペシフィシティーというの ...
0080: もう一方の特徴で(F あの)目的の ...
0081: (F えー)検索するということに関 ...
0082: <雑音>
0083: (F え)精度
0084: (F えー)プリシジョンと呼ばれて ...
0085: に関係するもんですけれど(F ...
0086: 内容
0087: を特徴的な要素を掴まえている
0088: という
0089: ことが(F ま)望ましい訳です
0090: で当然のことな ...

Center IPU of
Retrieved Passage

This relevant passage as a whole
is considered to be retrieved.

Fractional MAP (fMAP)



0072: <雑音>
0073: (D ぷそ)
0074: (F えー)漏れなくという方に関係し ...
0075: で(F その)評価尺度としていわゆる ...
0076: (F えー)検索ができているかという ...
0077: <雑音>
0078: (F え)もう一つの
0079: (F え)スペシフィシティーというの ...
0080: もう一方の特徴で(F あの)目的の ...
0081: (F えー)検索するということに関 ...
0082: <雑音>
0083: (F え)精度
0084: (F えー)プリシジョンと呼ばれて ...
0085: に関係するもんですけれど(F ...
0086: 内容
0087: を特徴的な要素を掴まえている
0088: という
0089: ことが(F ま)望ましい訳です
0090: で当然のことな ...

$$fAveP_q = \frac{1}{|R_q|}\sum_{i=1}^{|P_q|} \boxed{rel(p_i, R_q)} \frac{\sum_{j=1}^{i} \boxed{prec(p_j, R_q)}}{i}$$

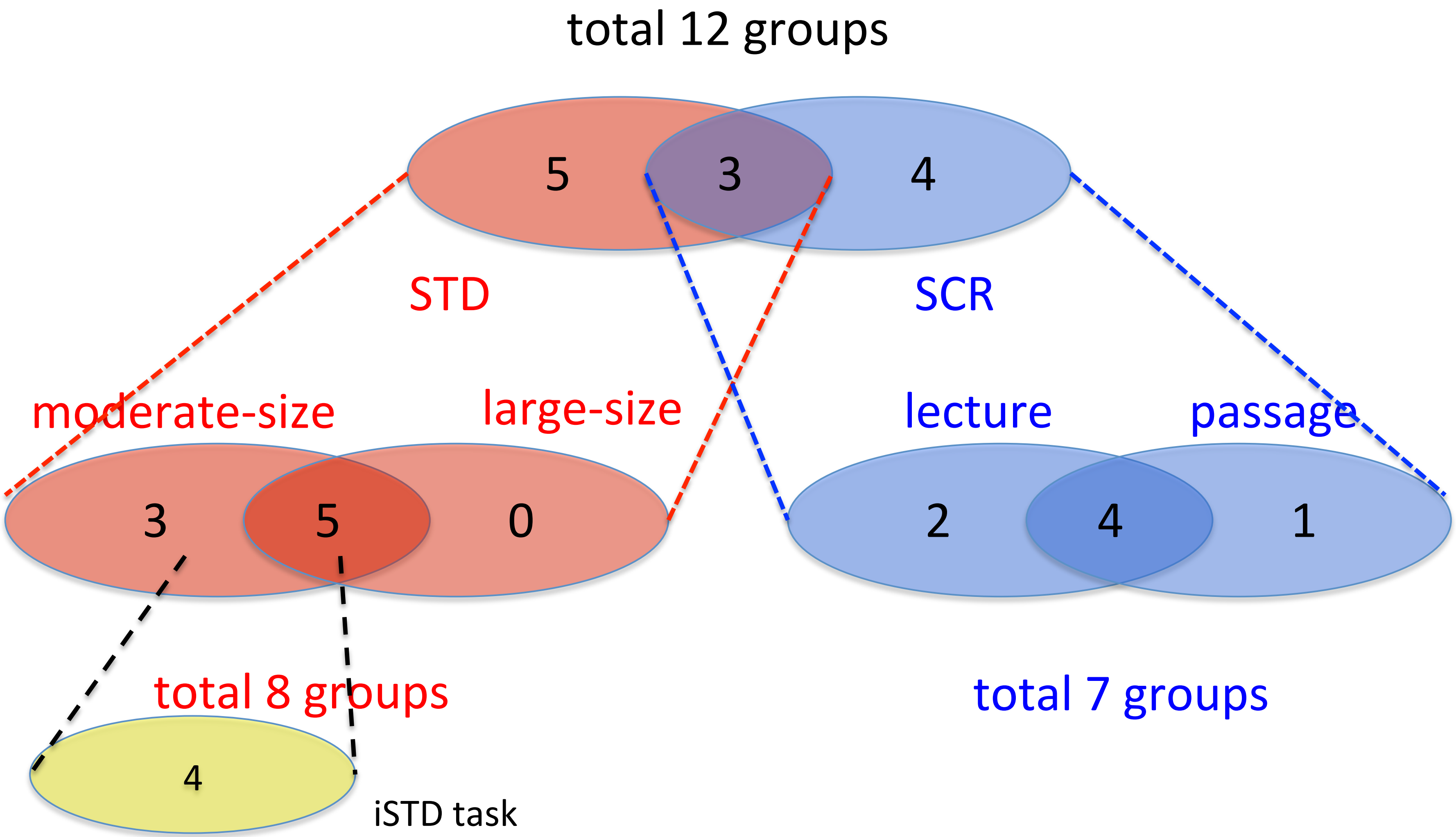
$$\boxed{prec(p, R_q) = \max_{r \in R_q} \frac{\boxed{|r \cap p|}}{\boxed{|p|}}}$$

$$\boxed{rel(p, R_q) = \max_{r \in R_q} \frac{\boxed{|r \cap p|}}{\boxed{|r|}}}$$

Outline

- ✓ Background
- ✓ Task Definition
  - ✓ Documents & Transcriptions
  - ✓ Subtasks
- Evaluation Results
  - STD task
  - SCR task

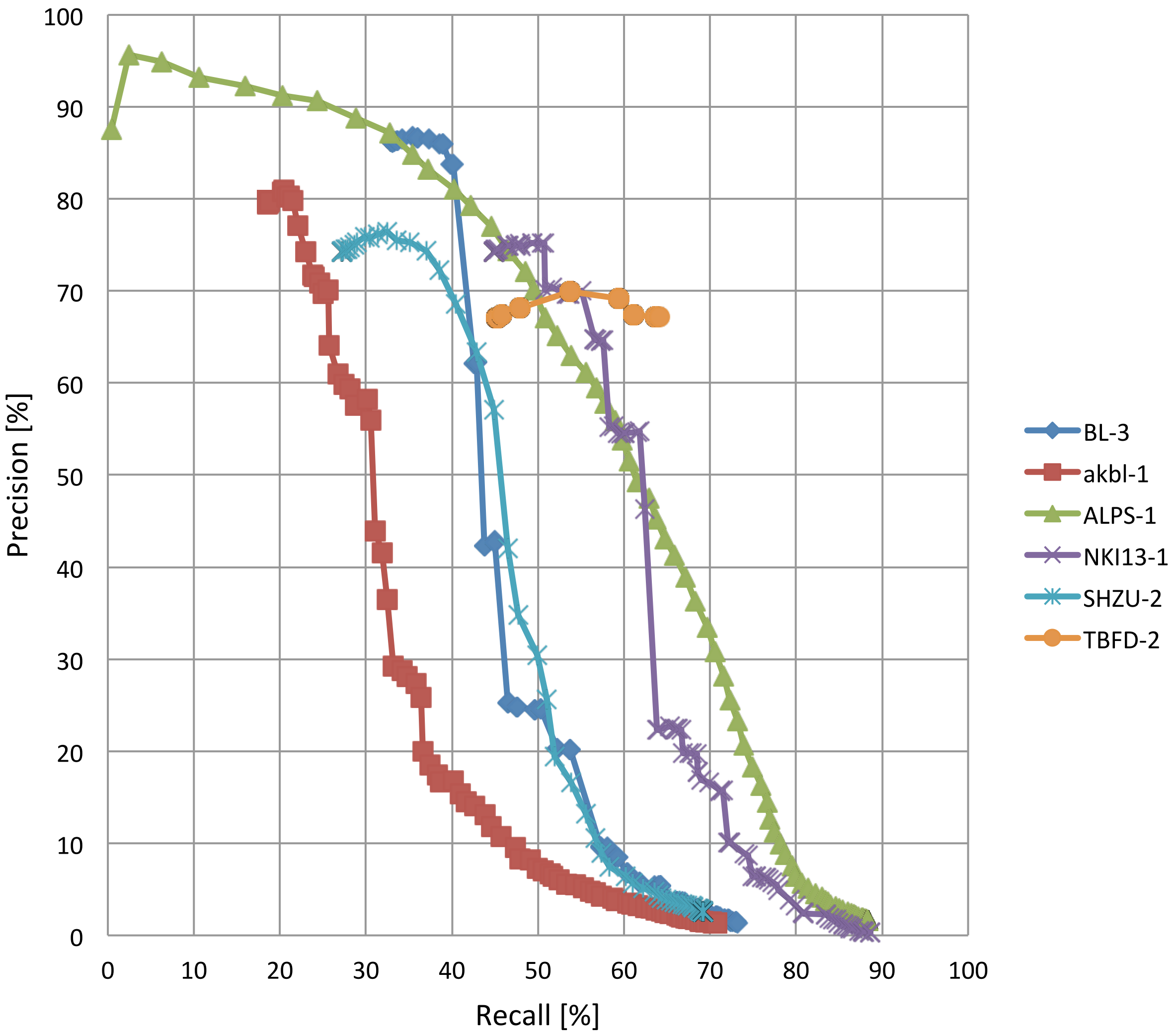
Participant Groups



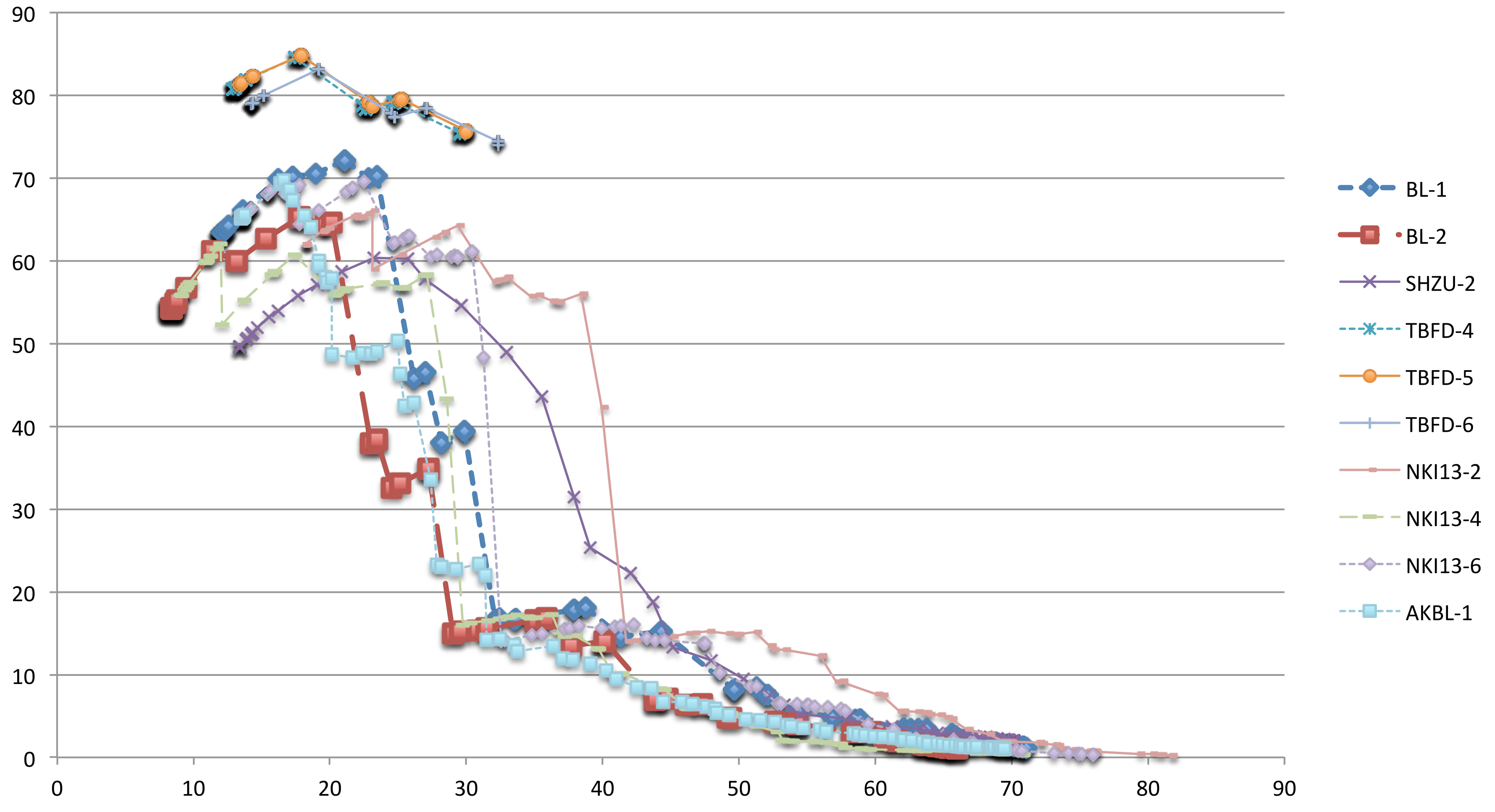
Baseline runs for STD task

- Edit distance-based Continuous DP matching against ...
  - BL-1: REF-SYLLABLE-MATCHED
  - BL-2: REF-WORD-MATCHED
  - BL-3: REF-WORD-MATCHED for IV query terms, REF-SYLLABLE-MATCHED for OOV query terms
- For iSTD task,
  - Rank query terms according to the lowest detection score throughout document collection.

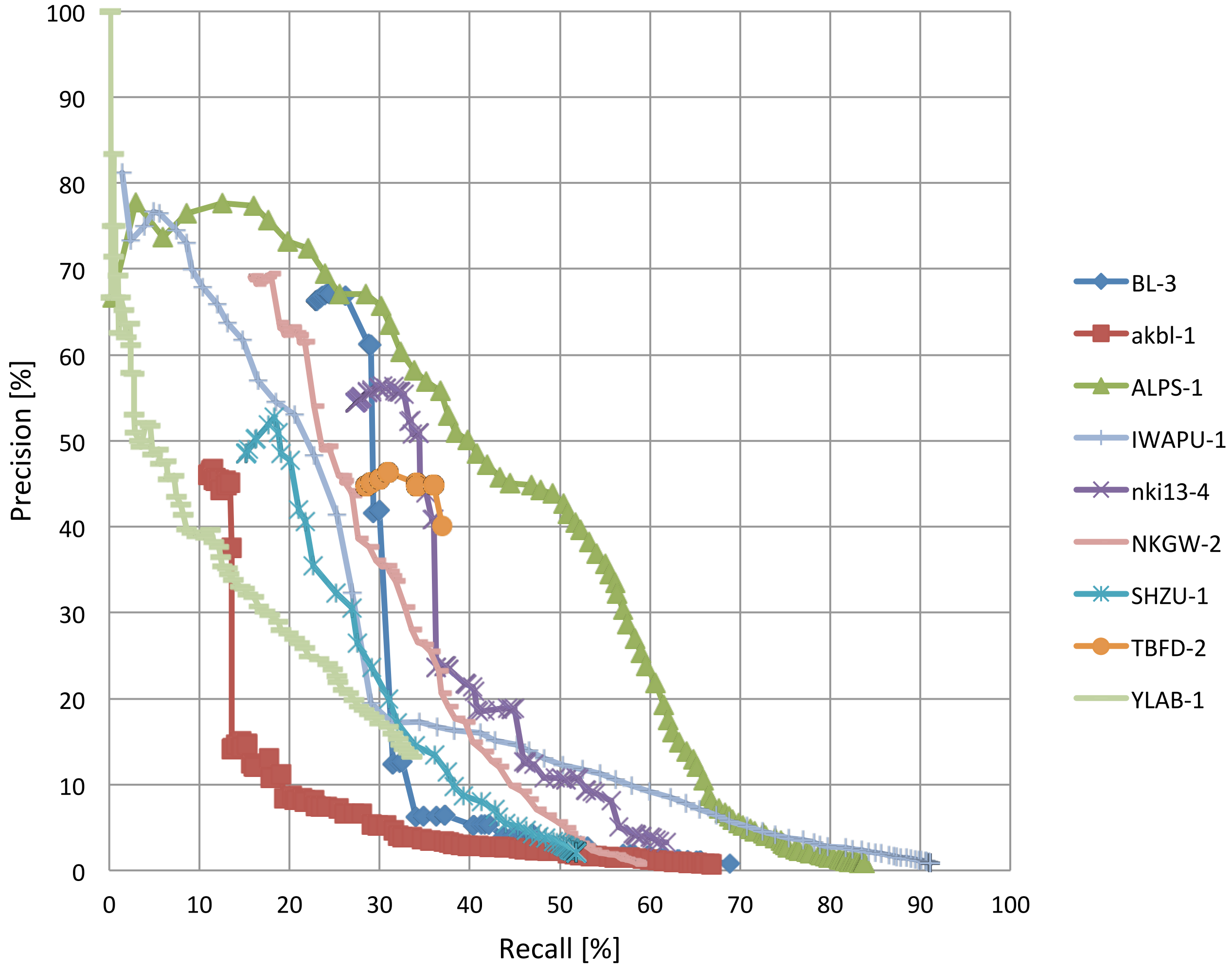
STD large-size task results



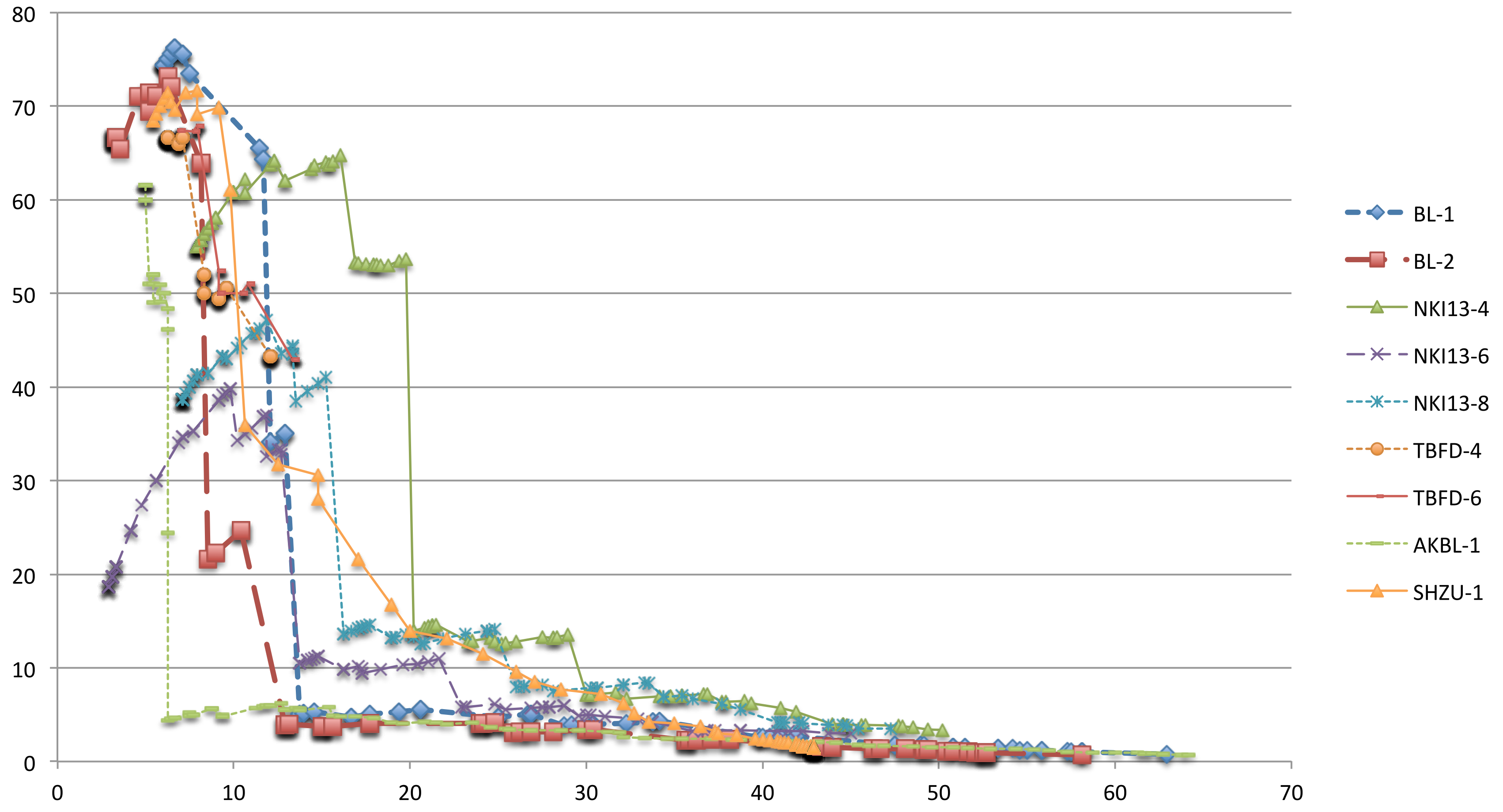
OOV term detection results at CSJ large-size task for the runs using only the reference MATCHED transcriptions



STD moderate-size task results



OOV term detection results at SDPWS moderate-size task for the runs using only the reference MATCHED transcriptions



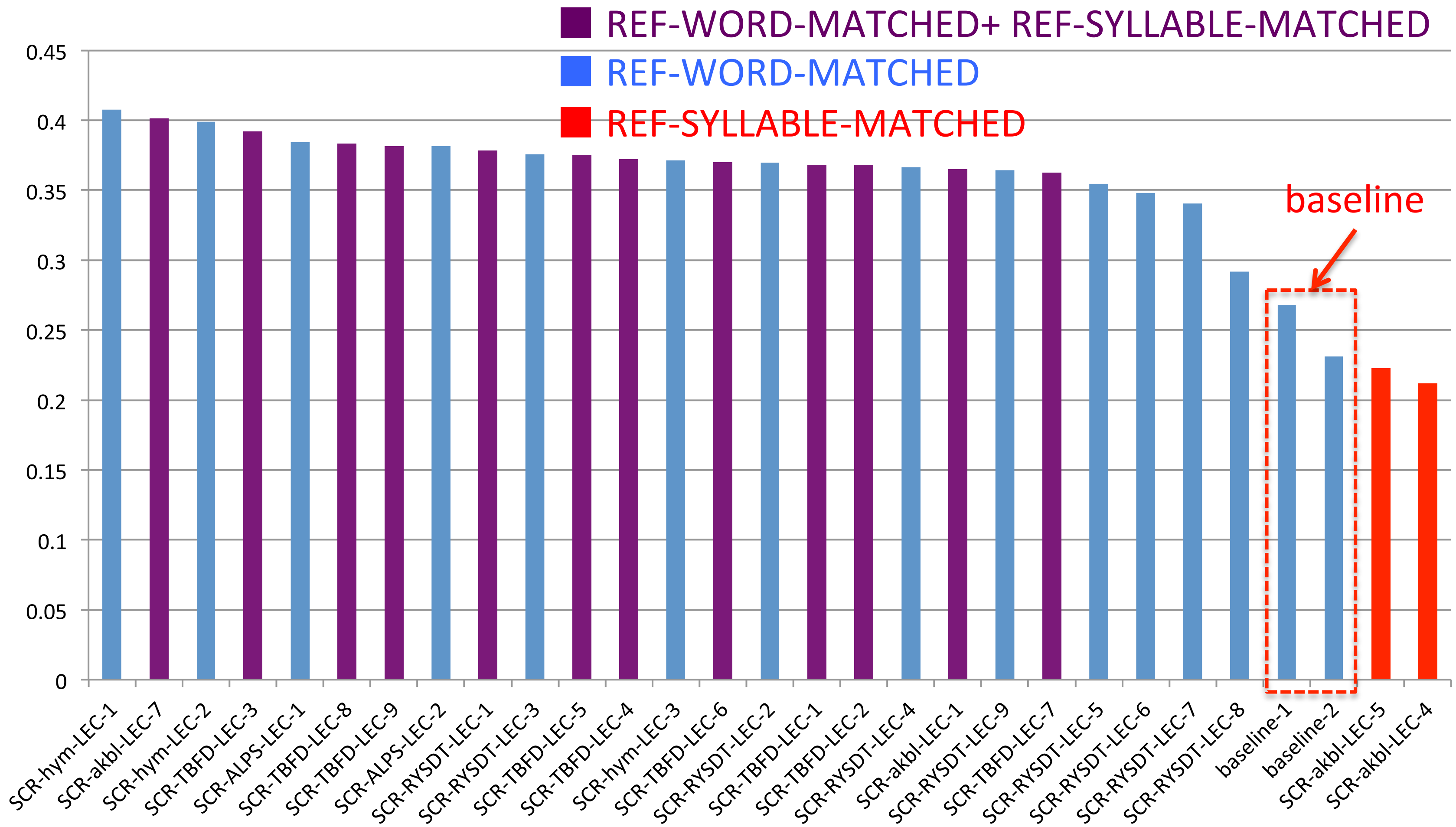
Baseline runs for SCR task

- Conventional word-based vector space model using 1-best transcription.
  - Transcription: REF-WORD-MATCHED or REF-WORD-UNMATCHED
  - Term weighting: TF-IDF with pivoted normalization (SMART) or without it (TF-IDF)

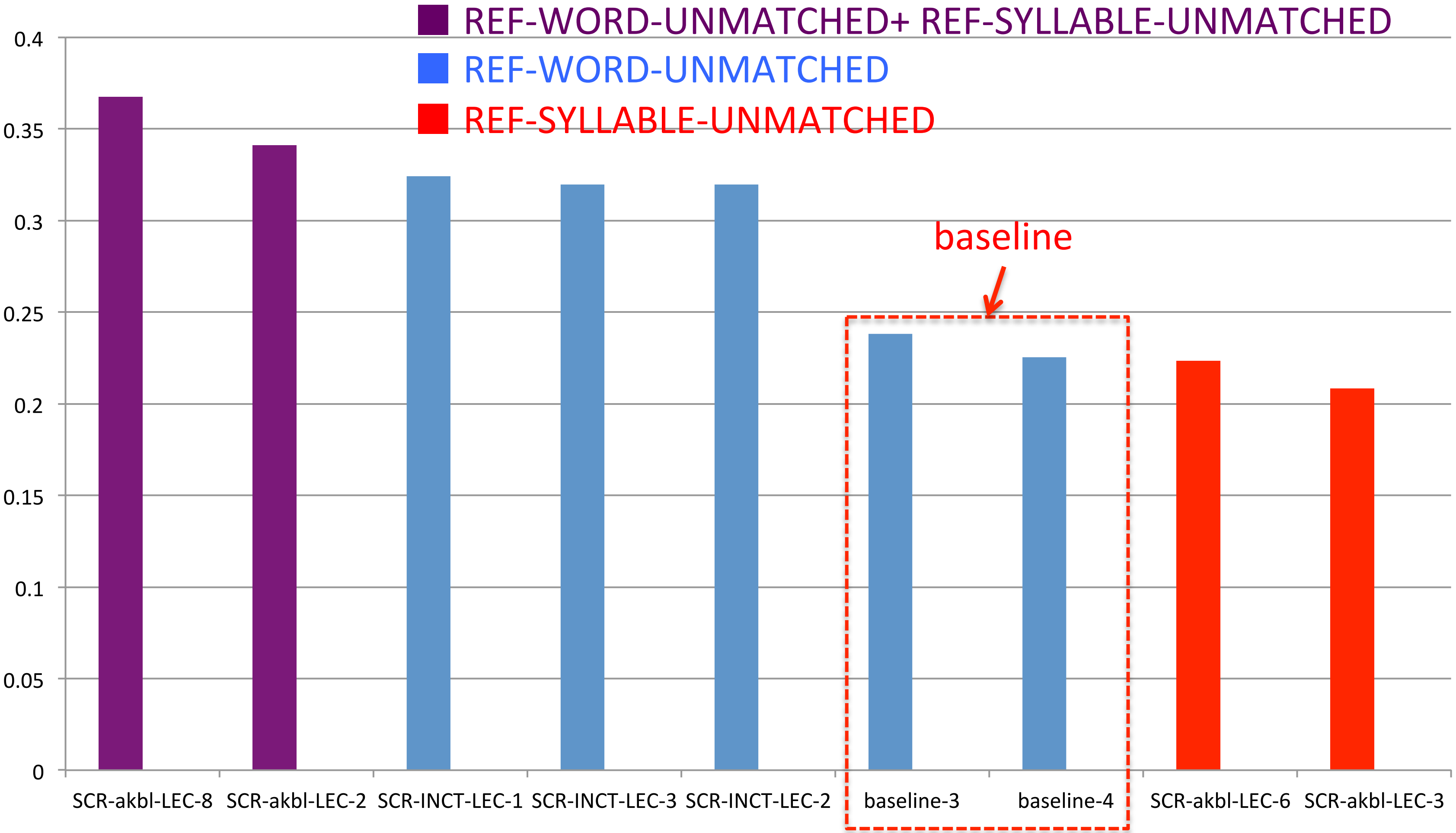
	REF-WORD-MATCHED	REF-WORD-UNMATCHED
SMART	baseline-1	baseline-3
TF-IDF	baseline-2	baseline-4

- For passage retrieval task,
  - Using pre-defined fixed length 15 IPUs as a passage

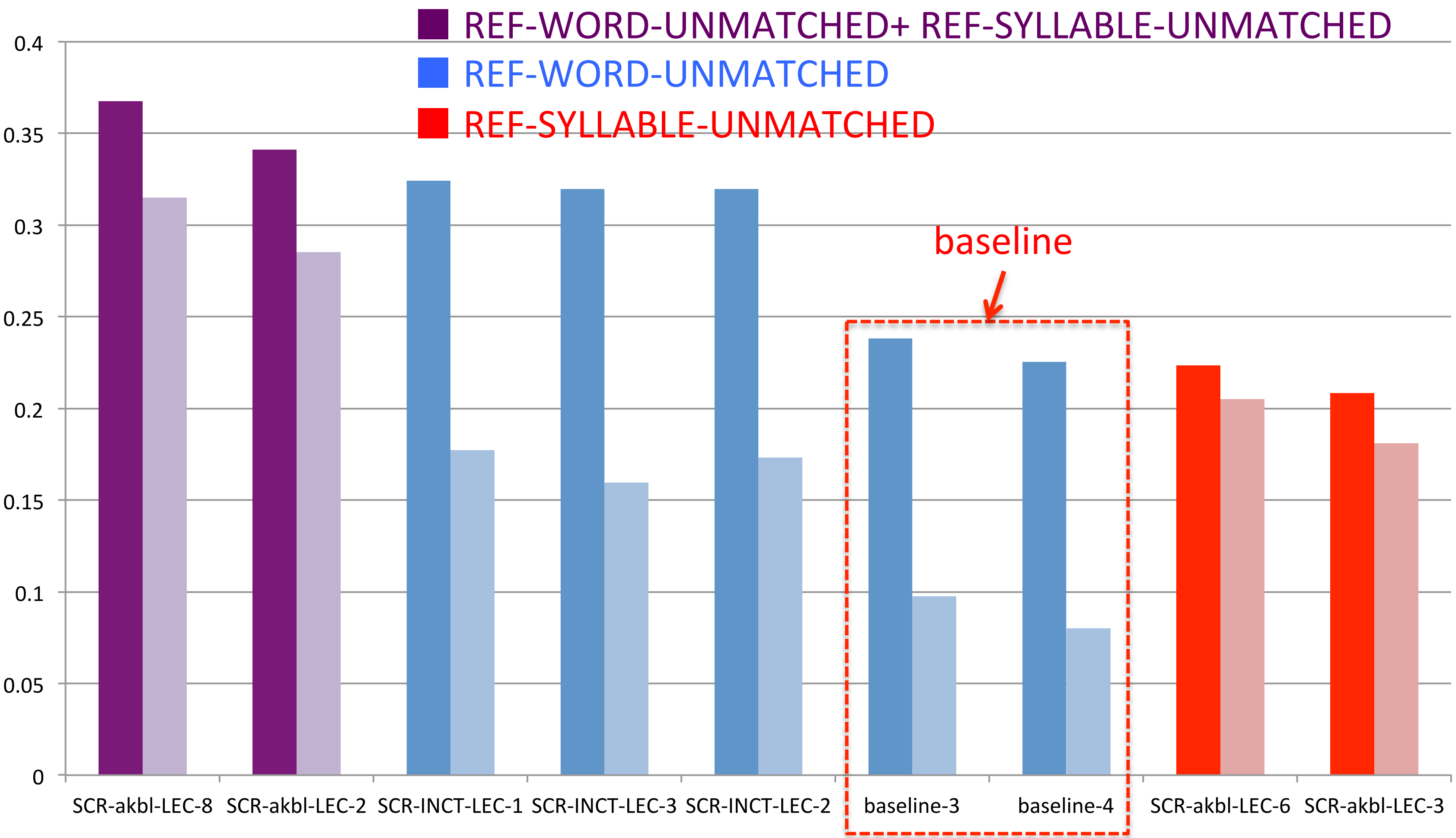
Lecture retrieval task result
(the runs using "matched" transcriptions)



Lecture retrieval task results
(the runs using "unmatched" transcriptions)

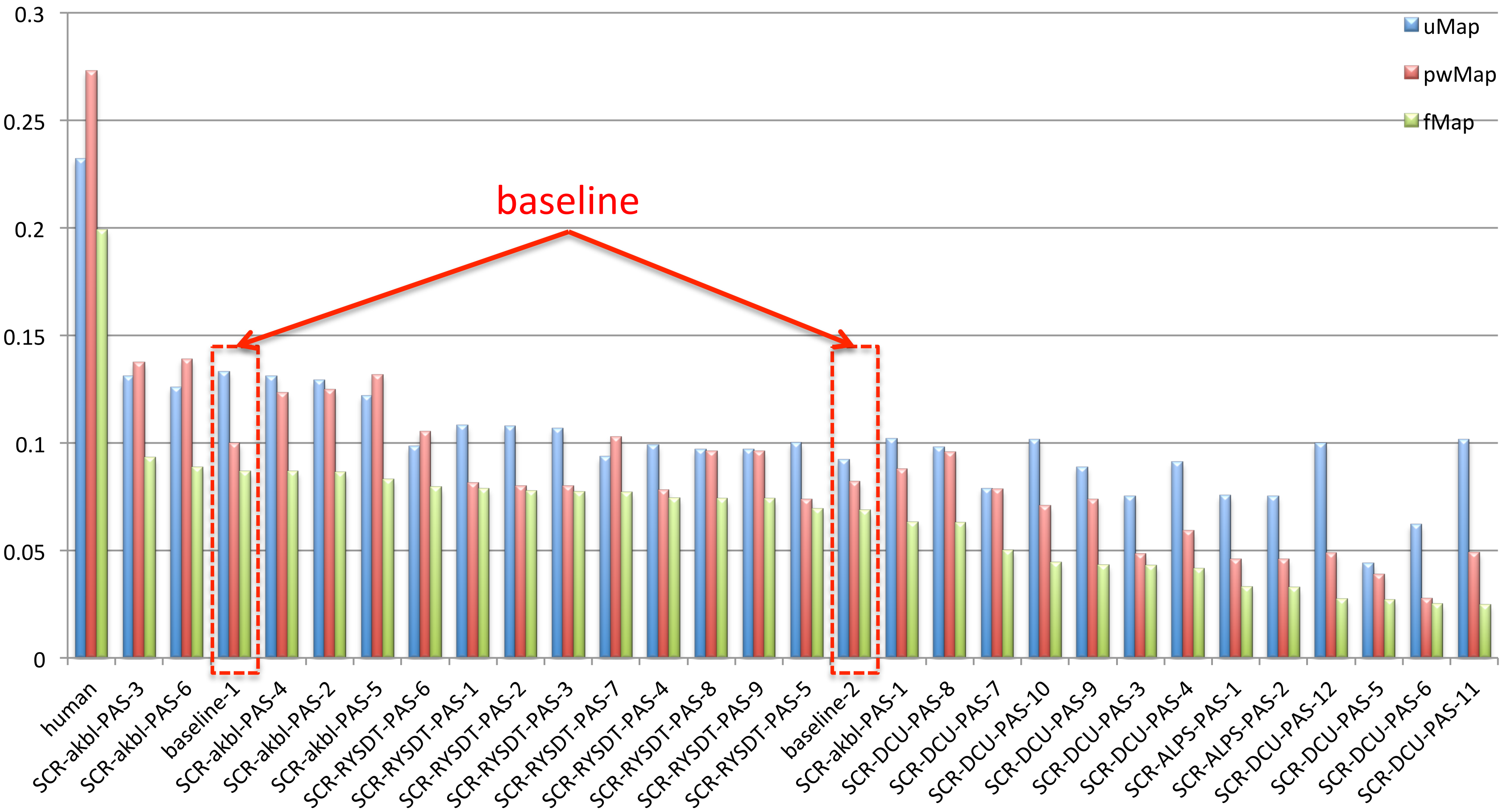


Lecture retrieval task results (the runs using "unmatched" transcriptions)



left: All query / right: OOV query

Passage retrieval task results (the runs using "matched" transcriptions)



Conclusion

- Findings for STD task,
  - Because STD task is OOV-sensitive, the comparison is difficult among runs using different automatic transcriptions.
  - For the OOV term detection, syllable-based transcription is more effective than word-based.
  - Along with the finding in the SpokenDoc-1, using multiple transcriptions (e.g. syllable and word transcriptions) is still effective, even when it does not work to reduce the OOV rate.
  - While variety of indexing methods are introduced in the SpokenDoc-2 and successfully improve their efficiency, their detection performance seem to depend on their underlying DTW-matching methods.
- Findings for SCR task,
  - Using both word-based and syllable-based transcription is effective, but, unlike STD, only for OOV queries.
  - Human performed much better than the current automated system for the boundary-free passage retrieval task, which indicates that there are still rooms for improvement.