SpokenDoc

IR (Information Retrieval) for Spoken Documents

Kiyoaki Aikawa (Tokyo University of Technology)
Tomoyosi Akiba (Toyohashi University of Technology)
Tatsuya Kawahara (Kyoto University)
Tomoko Matsui (The Institute of Statistical Mathematics)
Seiichi Nakagawa (Toyohashi University of Technology)
Hiroaki Nanjo (Ryukoku University)
Hiromitsu Nishizaki (Yamanashi University)
Hu Xinhui (NICT)
Yoichi Yamashita (Ritsumeikan University)

What is “SpokenDoc”?

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

Speech data ≅ Transcribed text by automatic speech recognition

“information retrieval”

ASR

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

Background

- 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 SDR task
  - First evaluation targeting Japanese and lecture speech
  - Investigate Boundary-free passage retrieval task

Outline

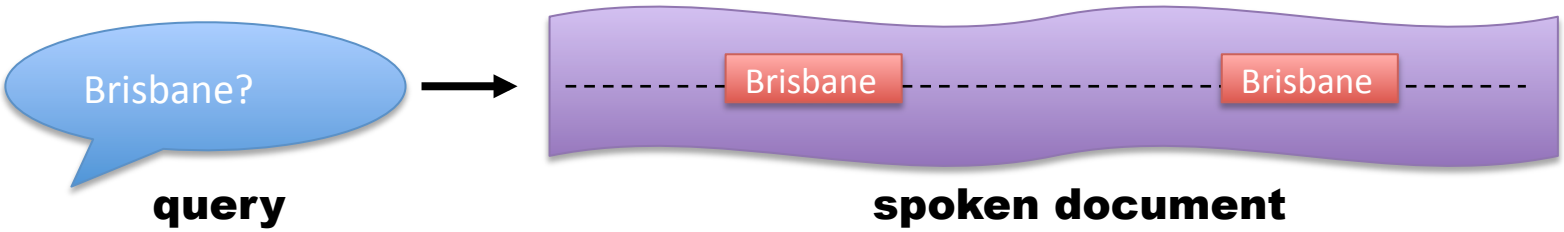
- ✓ Background
- Task Definition
  - Document Collection & Transcriptions
  - Subtasks
- Evaluation Results
  - STD subtask
  - SDR subtask
- SpokenDoc-2

文書データ

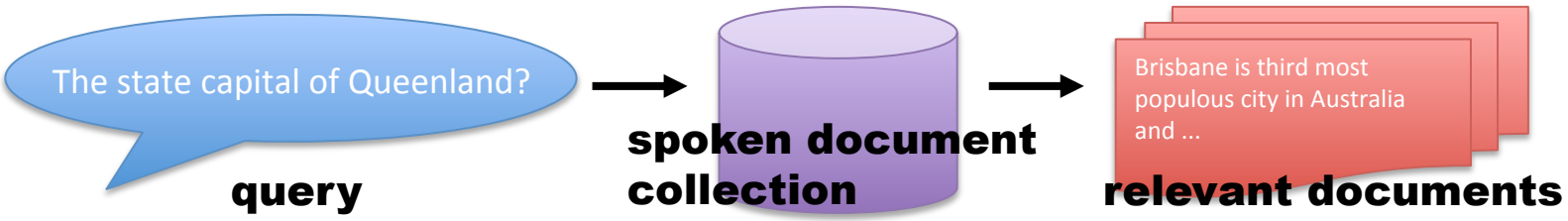
- 「日本語話し言葉コーパス(CSJ)」の学会講演と模擬講演 (628時間)
  - 国立国語研究所から入手する
- オーガナイザから2種類の音声認識結果を提供
  - 単語ベース音声認識結果
    - 27,000単語の単語辞書
    - 単語3-gram言語モデル
  - 音節ベース音声認識結果
    - 音節辞書
    - 音節3-gram言語モデル
  - 複数の認識候補表現
    - N-bestリスト, コンフュージョンネットワーク, ラティス

SpokenDocにおける
2つのサブタスク

- STD: Spoken Term Detection (音声検索語検出)

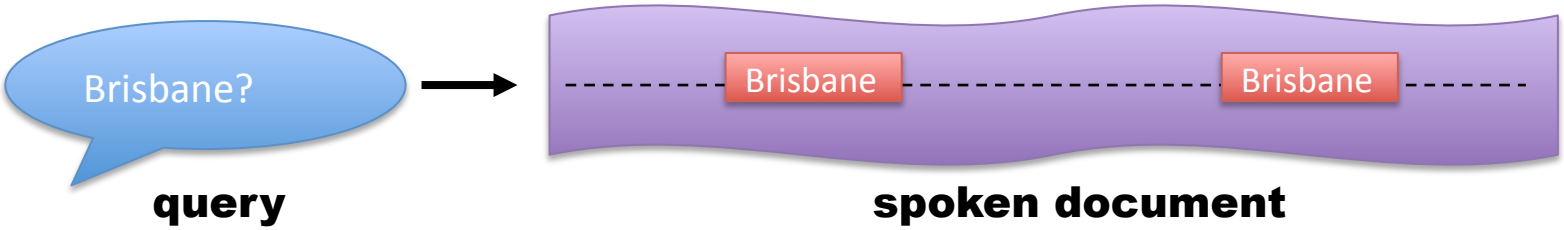


- SDR: Spoken Document Retrieval (音声内容検索)



STD サブタスク

- Spoken Term Detection (音声検索語検出)
  - 検索クエリ(単語列)がそのまま現れる部分を特定するタスク。
  - 探したいものを知っている(既知である)状況を想定。
- 検索対象:
  - ALL: 全2702講演
  - CORE: 177講演



STDサブタスクの評価指標

- 発話(IPU)を単位として求めた精度(Precision)および再現率(Recall)を基本とする。

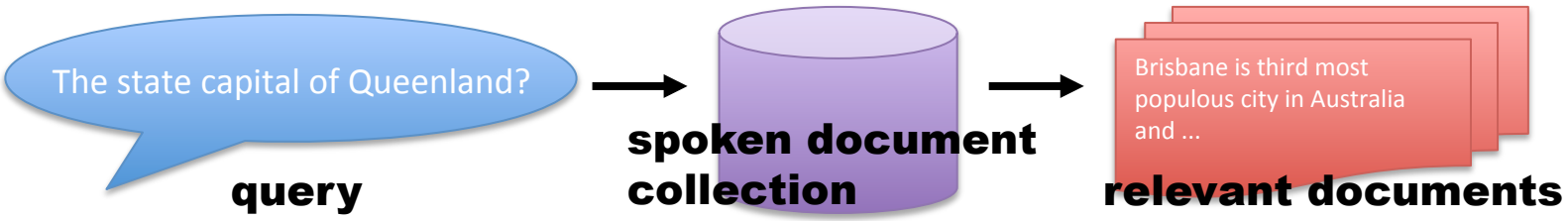
$$\text{Precision} = \frac{\text{検出したIPU集合} \cap \text{正解IPU集合}}{\text{検出したIPU集合}}$$

$$\text{Recall} = \frac{\text{検出したIPU集合} \cap \text{正解IPU集合}}{\text{正解IPU集合}}$$

- 評価尺度
  - 要約: Actual F-measure, Max F-measure, Mean Average Precision (MAP)
  - グラフ: Recall-Precision曲線

SDR サブタスク

- Spoken Document Retrieval (音声内容検索)
  - 検索クエリ(文)と内容が一致する部分を特定するタスク。
  - 探したいものが漠然としている状況を想定。
- 正解判定の単位
  - 講演検索タスク: 講演
  - パッセージ検索タスク: 講演の一部
    - 連続する発話(IPU)列
    - 長さは任意



パッセージ検索タスクの例

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)

正解パッセージ

正解パッセージ

A01M0958

SDRサブタスクの評価指標

- 講演検索タスク
  - Mean Average Precision (MAP)
- パッセージ検索タスク
  - Utterance-based metric
    - utterance-based MAP (uMAP)
  - Passage-based metric
    - point-wise MAP (pwMAP)
    - fractional MAP (fMAP)

パッセージ検索の評価

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。

システムの検索結果

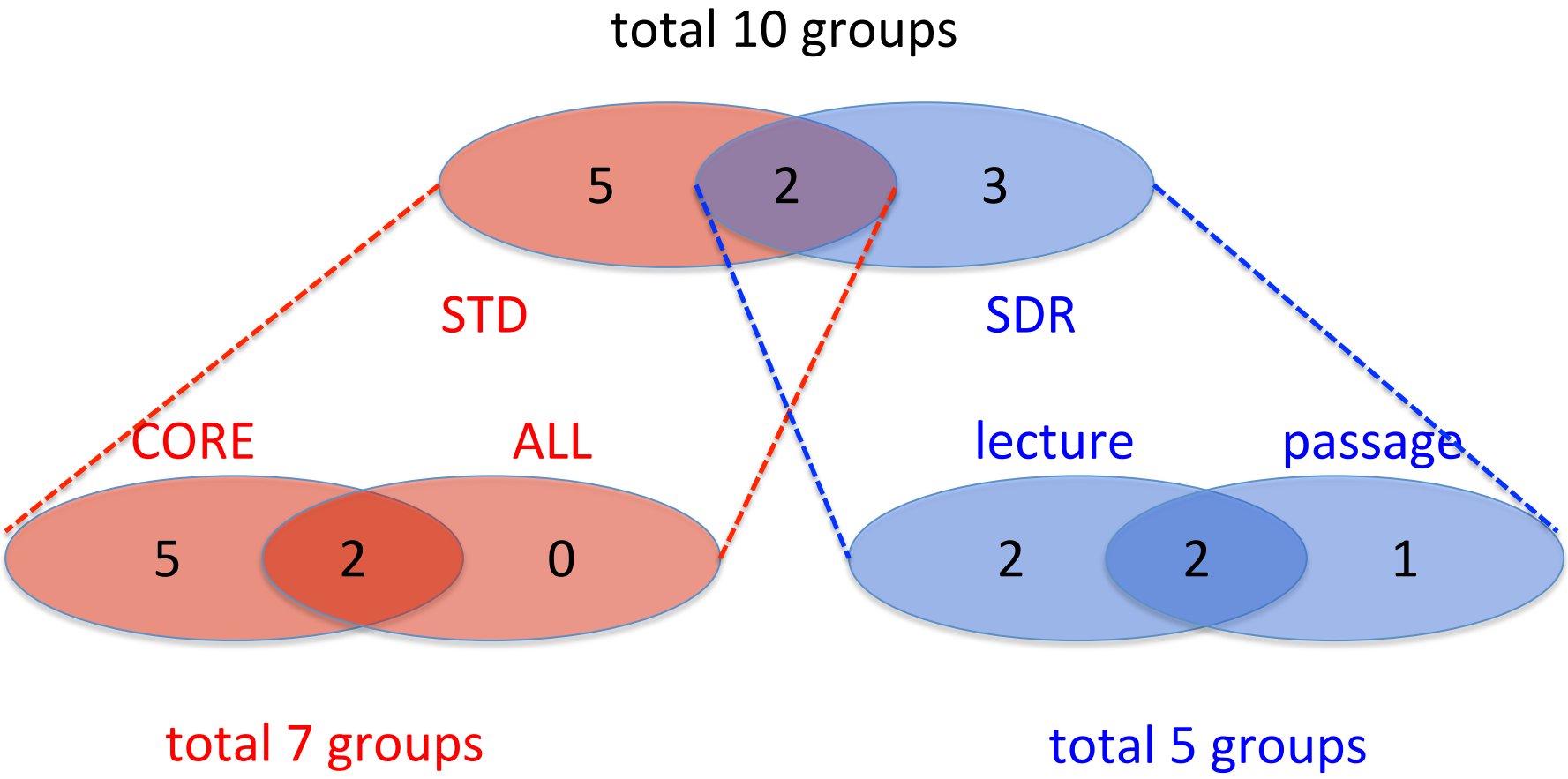
0072: <雑音>
0073: (D ぷそ)
0074: (F えー)漏れなくという方に関係している
0075: で(F その)評価尺度としていわゆる再現率と呼ばれているものは(F その)どれだけ(D も)網羅的に
0076: (F えー)検索ができているかということを表わす尺度です
0077: <雑音>
0078: (F え)もう一つの
0079: (F え)スペシフィシティーというのは(F そのー)
0080: もう一方の特徴で(F あの)目的の重要な要素である(F その)正確に
0081: (F えー)検索するということに関係してますこれは(F あのー)評価尺度で言うと
0082: <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

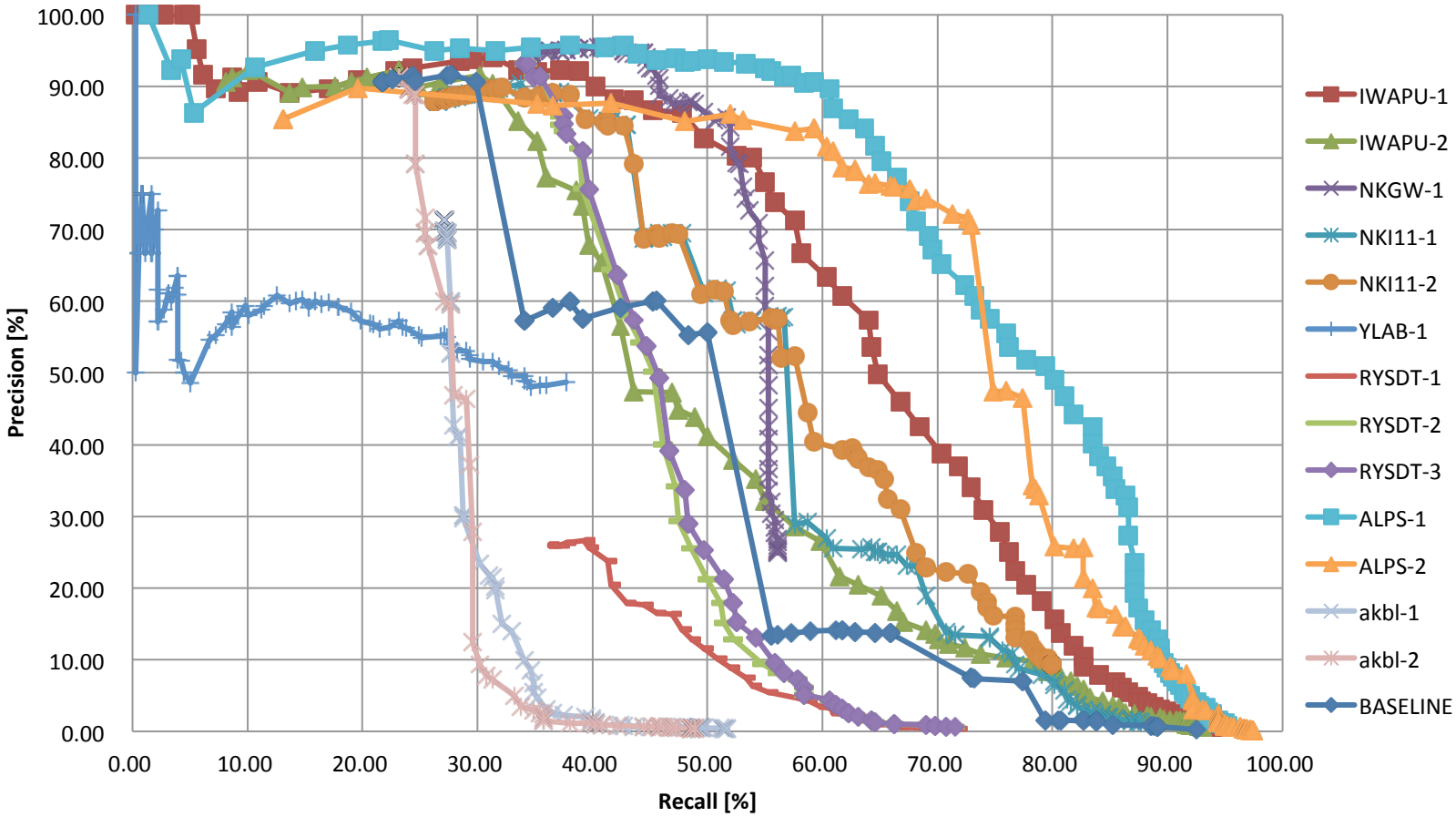
Outline

- ✓ Background
- ✓ Task Definition
  - ✓ Document Collection & Transcriptions
  - ✓ Subtasks
- Evaluation Results
  - STD subtask
  - SDR subtask
- SpokenDoc-2

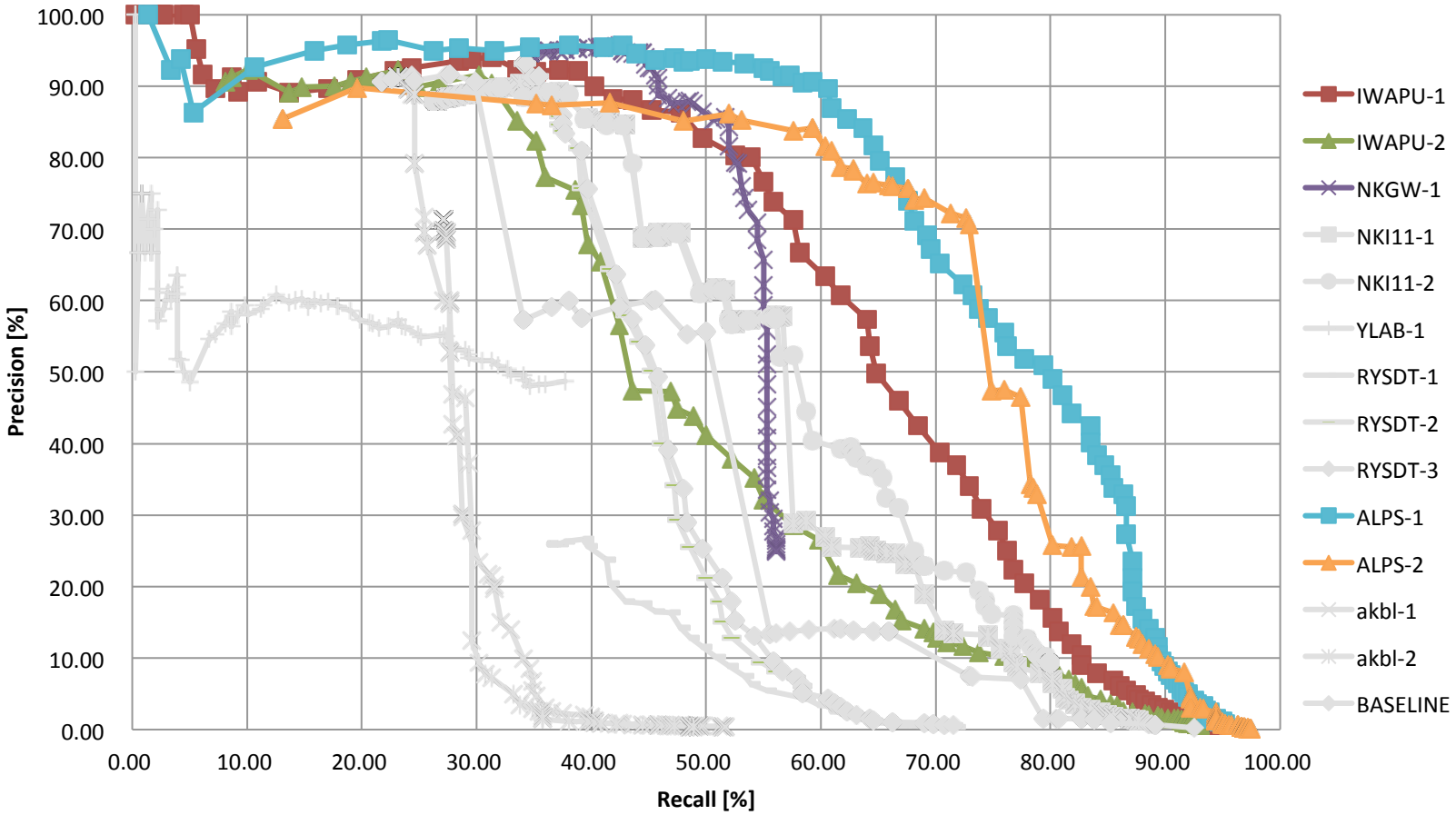
Participant Groups



Recall-Precision Curves for STD subtask for CORE set

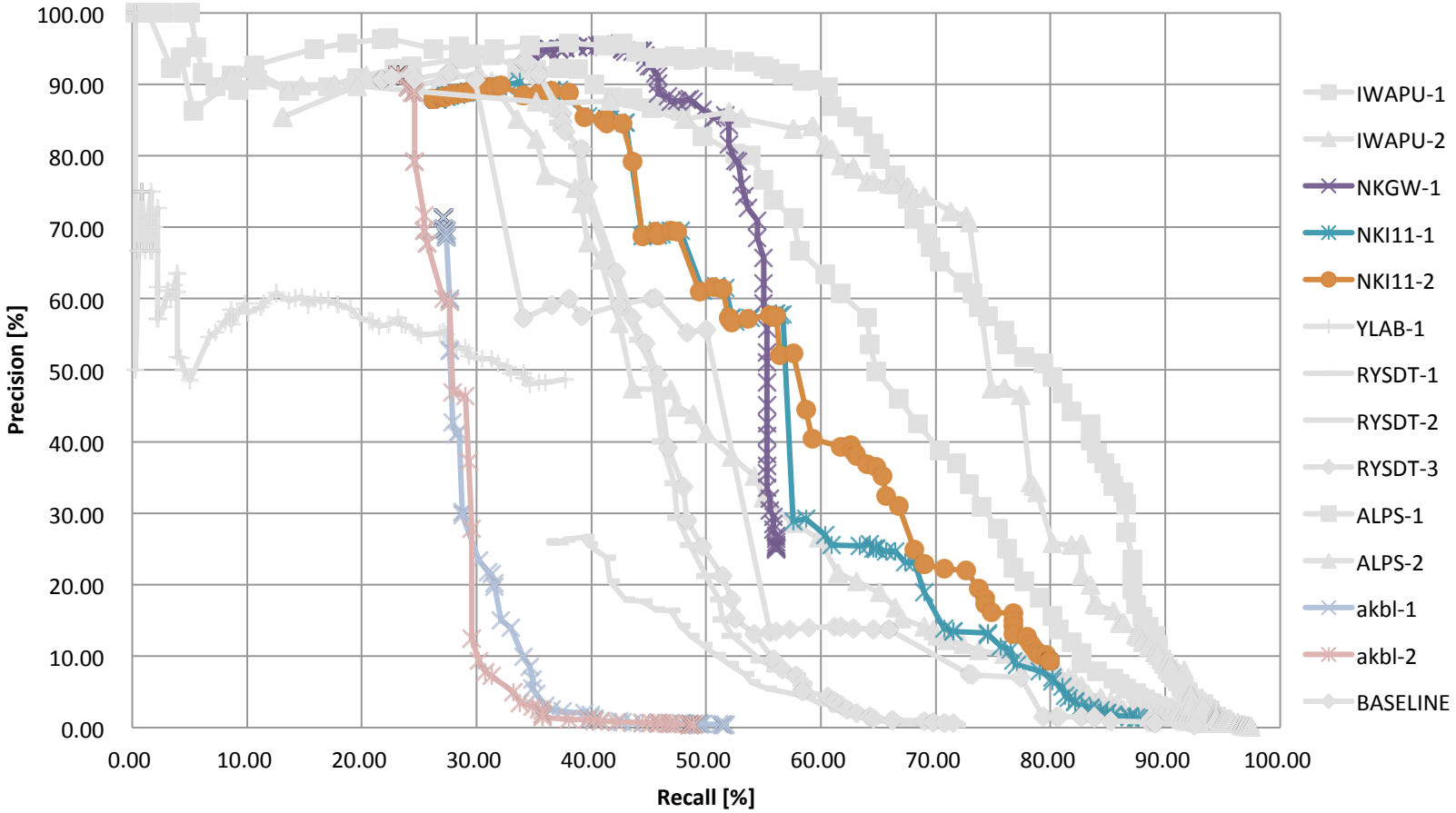


Recall-Precision Curves for STD subtask for CORE set



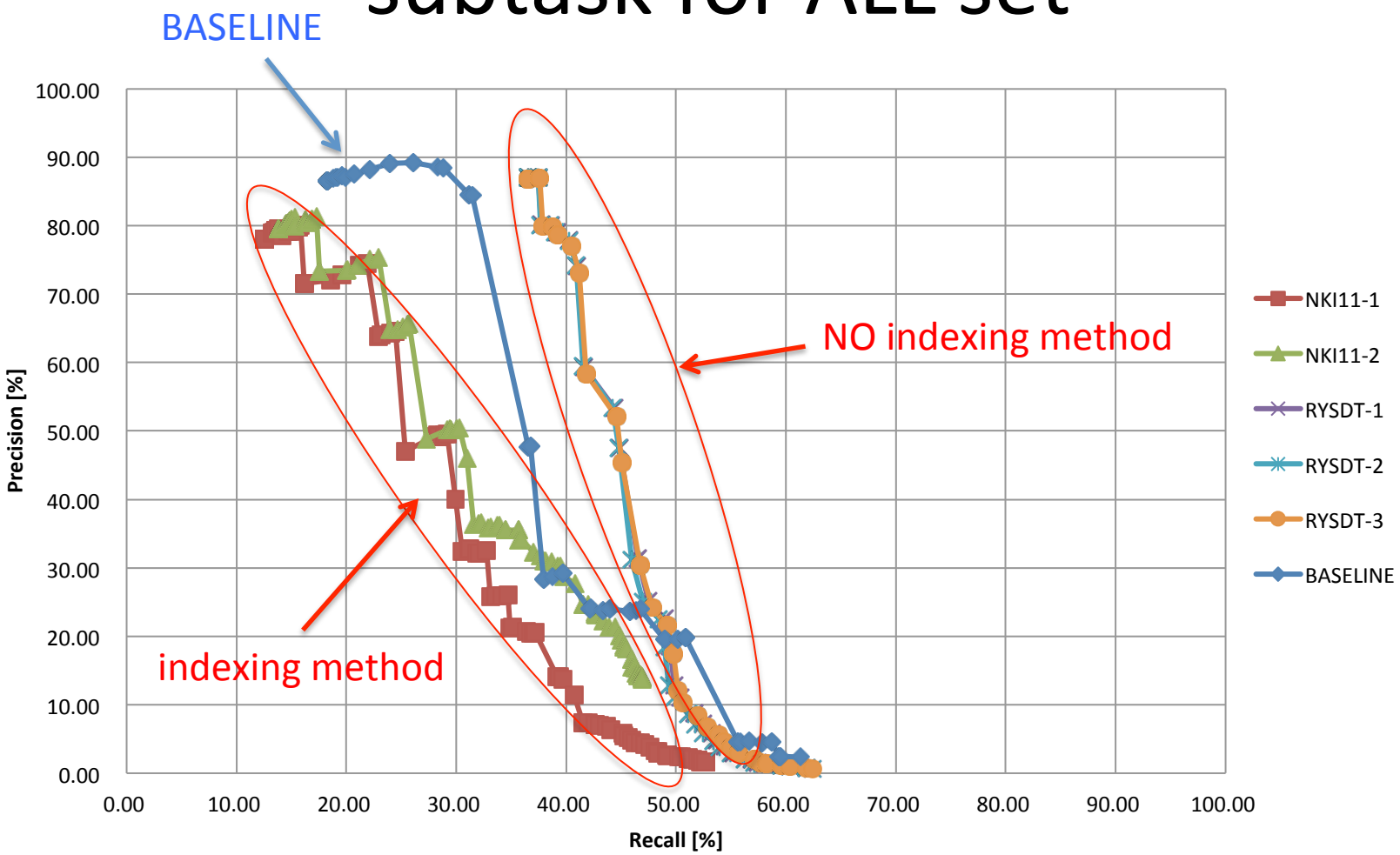
5 runs using multiple transcriptions

Recall-Precision Curves for STD subtask for CORE set

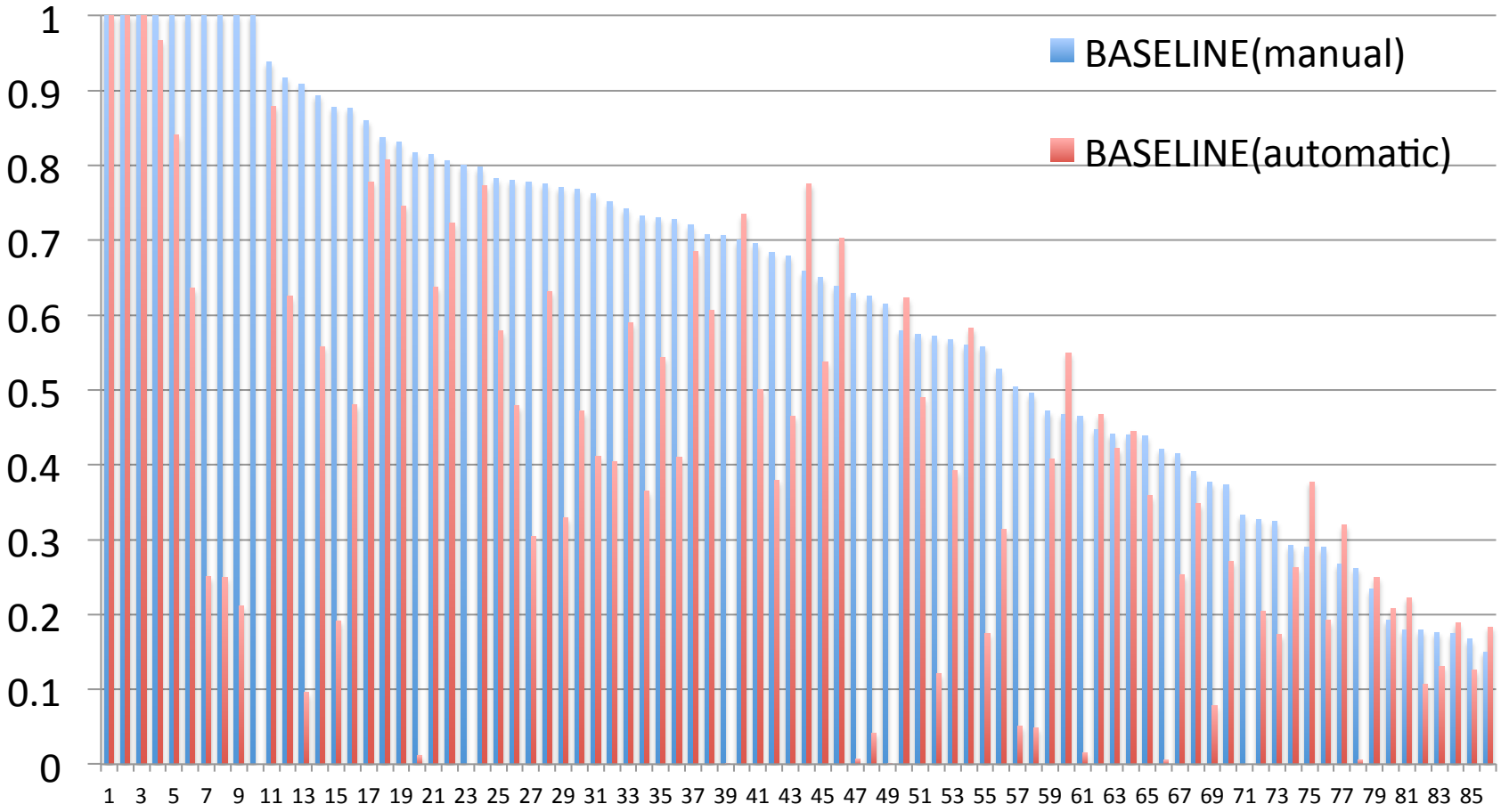


5 runs using indexing for efficiency

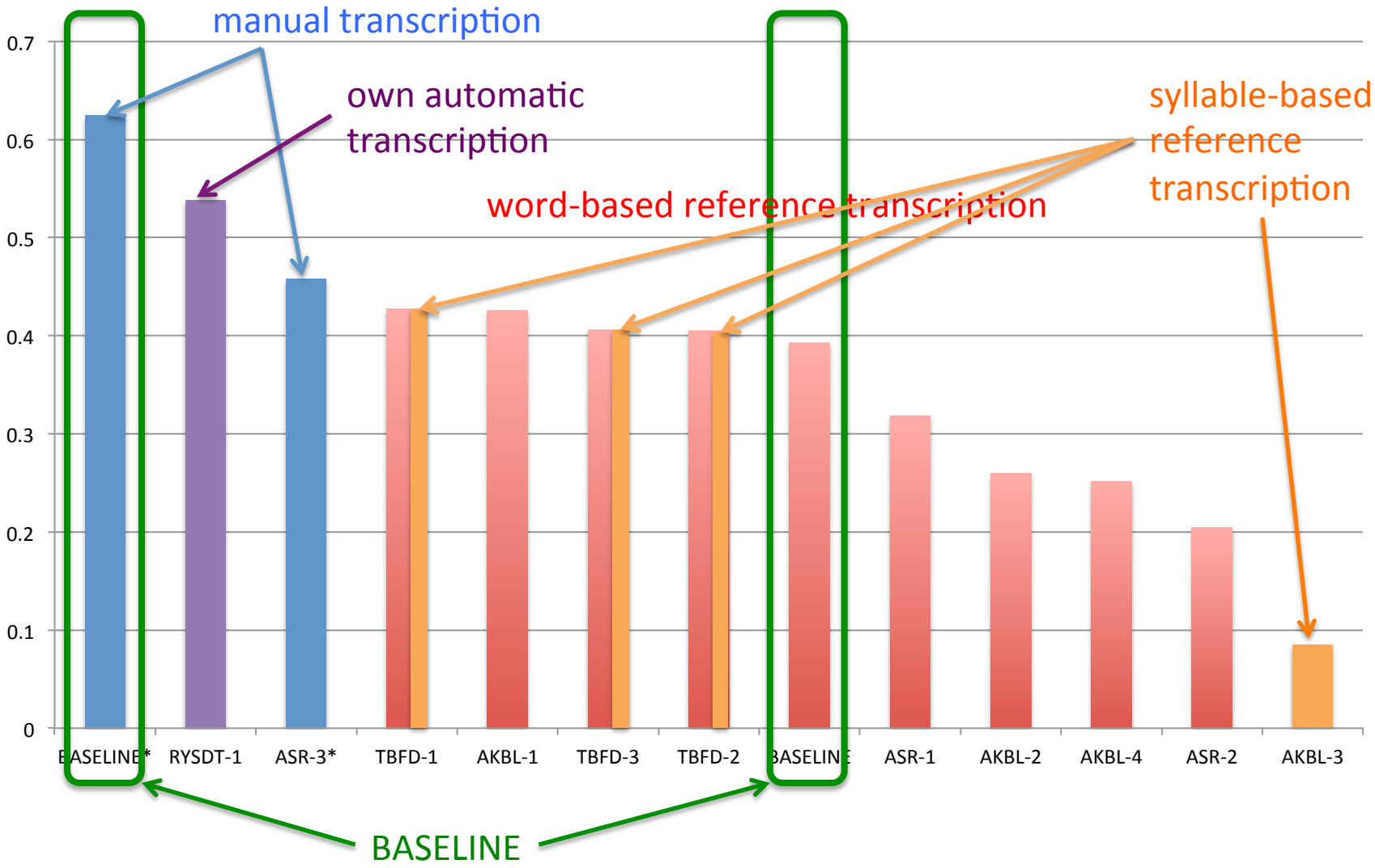
Recall-Precision Curves for STD subtask for ALL set



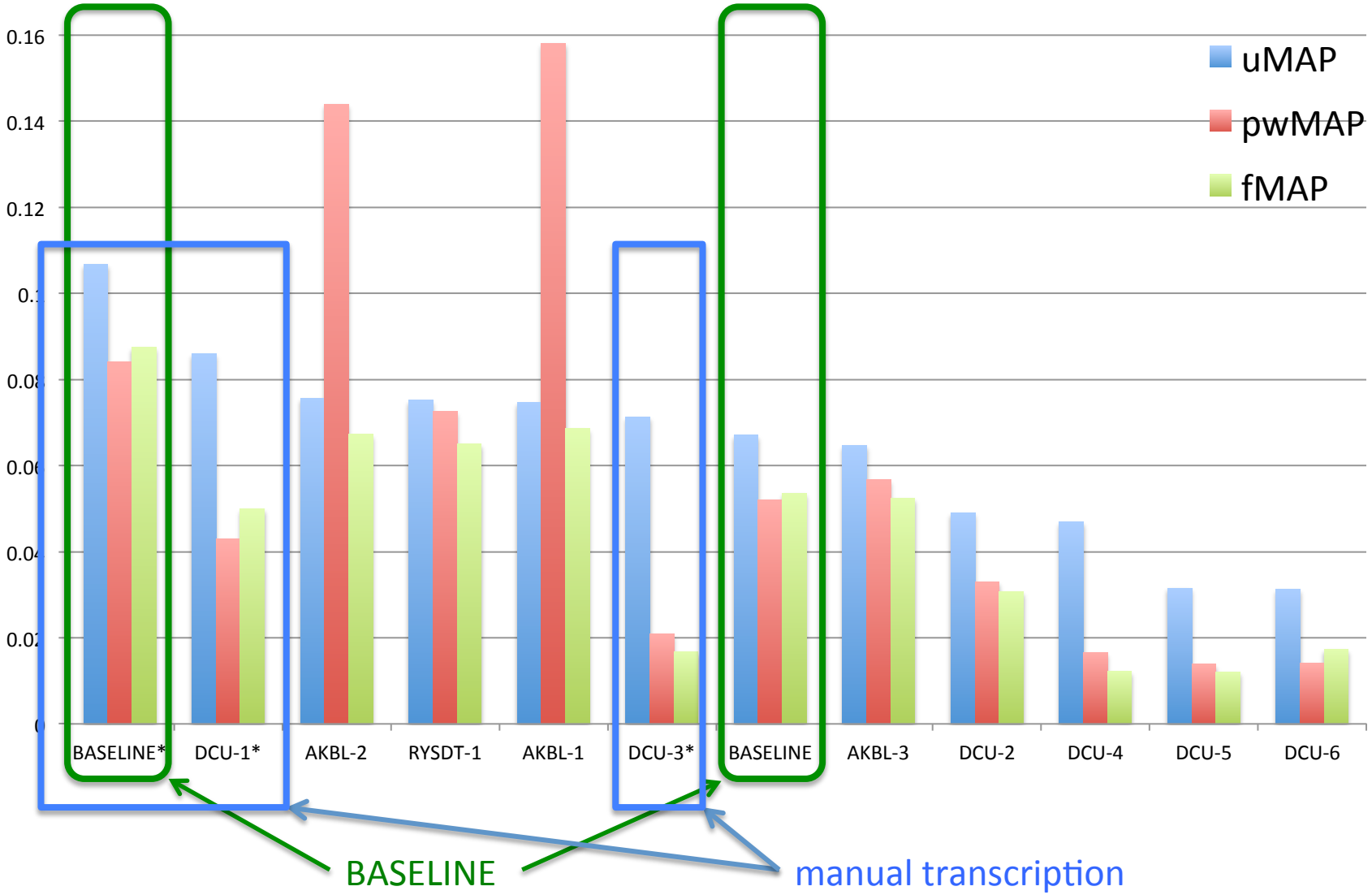
BASELINE Performance (MAP) per Query for SDR lecture retrieval task



Evaluation Result (MAP) for the lecture retrieval of SDR subtask



Evaluation Result for the passage retrieval of SDR subtask



Outline

- ✓ Background
- ✓ Task Definition
  - ✓ Document Collection & Transcriptions
  - ✓ Subtasks
- ✓ Evaluation Results
  - ✓ STD subtask
  - ✓ SDR subtask
- SpokenDoc-2

What's New in SpokenDoc-2?

- 文書データ
  - 日本語話し言葉コーパス(CSJ)
  - 第１回～第６回音声ドキュメント処理ワークショップでの学会講演(new)
    - CSJの入手が困難なチームも参加可能！
- 音声認識結果
  - マッチ条件の連続単語および連続音節認識結果
  - ミスマッチ条件の音声認識結果(new)
    - より現実的な条件での評価
- タスク
  - STDサブタスク: 語が発話されていないことを検出するタスク
  - SDRサブタスク: 可変長パッセージ検索タスク

We are welcome to join SpokenDoc-2

- Schedule
  - Mar. 2012: release of the task description
  - June 2012: release of the reference automatic transcriptions
  - Sept. 2012 dry-run
  - Nov. 2012: formal-run evaluation
  - Nov. 2012-Feb.2013: relevance judgment
  - Feb. 2013: formal-run evaluation results release
  - May 2013: camera ready submission due
  - June 2013: NTCIR-10 workshop meeting
- Contact
  - SpokenDoc-2 organizers:
    - E-mail: ntcadm-spokendoc2@cl.ics.tut.ac.jp
  - Web (coming soon!)
    - http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?ntcir10
  - Twitter: @spokendoc2