Engaging Content
Engaging People

DCU at the NTCIR-12 SpokenQuery&Doc-2 Task

David N. Racca Gareth J.F. Jones

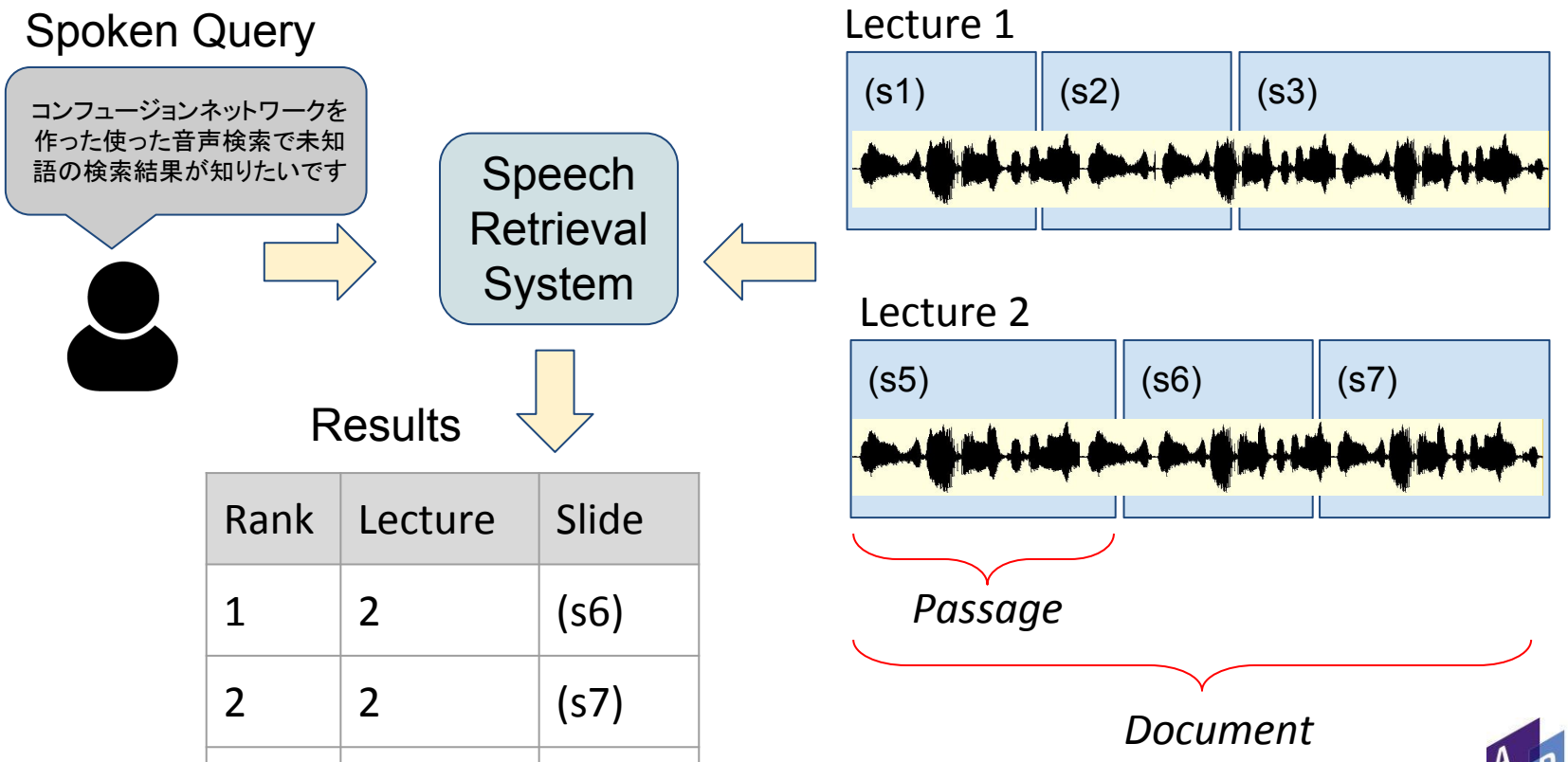
Dublin City University
Ireland

Outline

- Introduction
- Motivation and related work
- Data processing
- Retrieval models
- Parameter optimisation
- Experiments and results
- Conclusions

Introduction

- We participated in the *slide-group* retrieval task:
  - *"Find predefined spoken passages relevant to a spoken query"*



Results

Rank	Lecture	Slide
1	2	(s6)
2	2	(s7)
3	1	(s1)

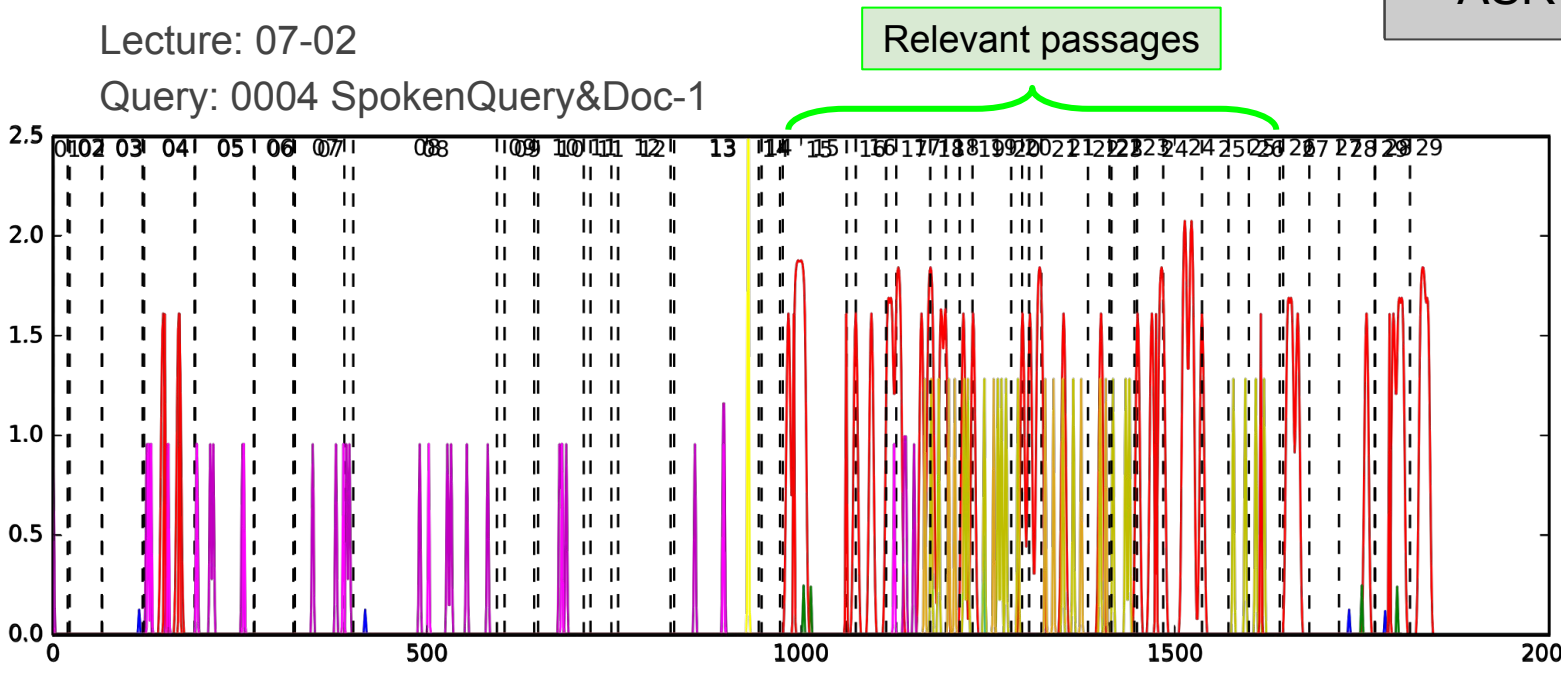
Introduction

- Document-based contextualisation (global)*:*
  - Considers a passage in the context of its document.
- Passage ranking implemented by combining passage and document retrieval scores computed with Okapi BM25.
- Query expansion (QE) with pseudo relevance feedback.
- BM25 and QE parameters optimised to improve the ranking quality of passages and documents.

Motivation

- "Short" passages as retrieval units:
  - are less redundant (i.e. contain fewer term repetitions);
  - are less likely to match query terms (i.e. less *"retrievable"*);
  - provide less reliable statistic estimates;
  - are more sensitive to ASR errors.

ASR case



Motivation

- Passages that share a document are likely to be about similar topics.
  - Passage relevance cannot be assumed independent:

$$D = p_1, \ldots, p_n \qquad R_i : p_i \text{ is relevant}$$

$$P(R_i) = 0.005 \ll 0.28 = P(R_i | R_{i-1})$$

Based on 237 queries and relevance data from:

- NTCIR-10 SpokenDoc
- NTCIR-11 SpokenQueryDoc-1
- NTCIR-12 SpokenQueryDoc-2

Related work

- Document-based contextualisation (global)*:*
  - Considers a passage in the context of its document.
- Neighbour-based contextualisation (local):
  - Considers neighbouring passages as context.
- Proven to be effective in previous editions of this task:
  - H. Nanjo et. al. "*Spoken document retrieval experiments for SpokenQuery&Doc at Ryukoku University (RYSDT)"*. In Proc. of NTCIR-11, 2014.
  - K. Hara et. al. "*Segmented spoken document retrieval using word co-occurrence information"*. In Proc. of NTCIR-11, 2014.
  - S. R. Shiang et. al. "*Spoken term detection and spoken content retrieval: Evaluations on NTCIR-11 SpokenQuery&Doc task"*. In Proc. of NTCIR-11, 2014.

Data processing

Spoken Lecture

ASR

ASR transcript

Untokenisation

0229: えーっと報告があるかと思う
0230: (s17)で実際に書く起こすの方
0231: (s18)行なうますたたえーと講義の収録を行なうて
0232: でこちらの方はえーとマイクの
0233: えーっと収録のする方やコンテンツが違うますので

Tokenisation, lemmatisation and POS filtering (MeCab)

0229: 報告 ある 思う
0230: (s17) 書く 起こす 方
0231: (s18) 行なう 講義 収録 行なう
0232: こちら 方 マイク
0233: 収録 する方 コンテンツ 違う

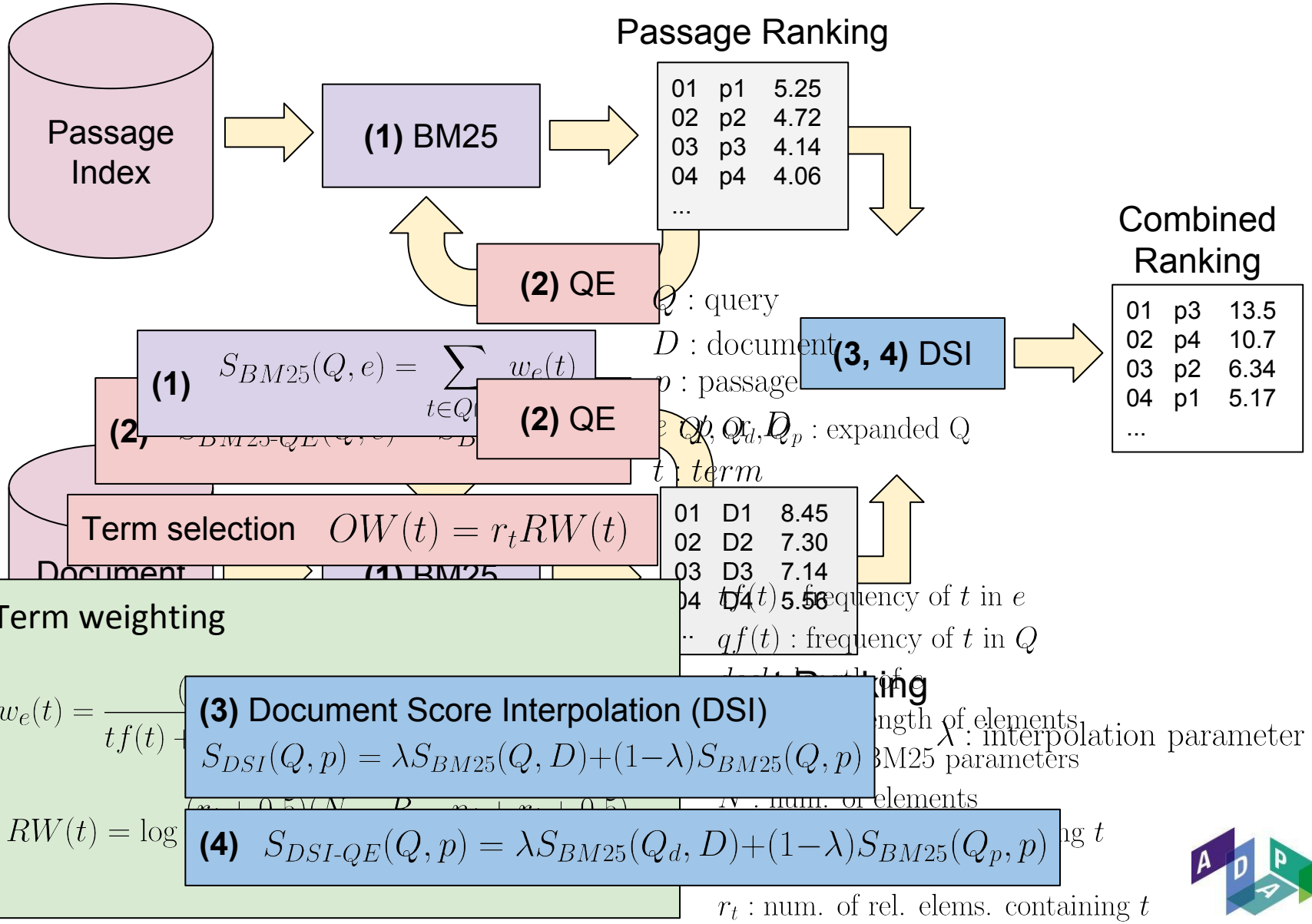
Stop words removal

0229: 報告 ある 思う
0230: (s17) 書く 起こす 方
0231: (s18) 行なう 講義 収録 行なう
0232: 方 マイク
0233: 収録 する方 コンテンツ 違う

Slide-group segmentation

0229: 報告 ある 思う
0230: **(s17)** 書く 起こす 方
0231: **(s18)** 行なう 講義 収録 行なう
0232: 方 マイク
0233: 収録 する方 コンテンツ 違う

Retrieval models



Retrieval models

- The SDPWS collection is rather *small*:
  - 98 lectures.
  - ~2300 passages (slide-groups).
  - ~340000 tokens (nouns & verbs only)
- Terms that would otherwise appear frequently in a large collection are underrepresented in the SDPWS collection.

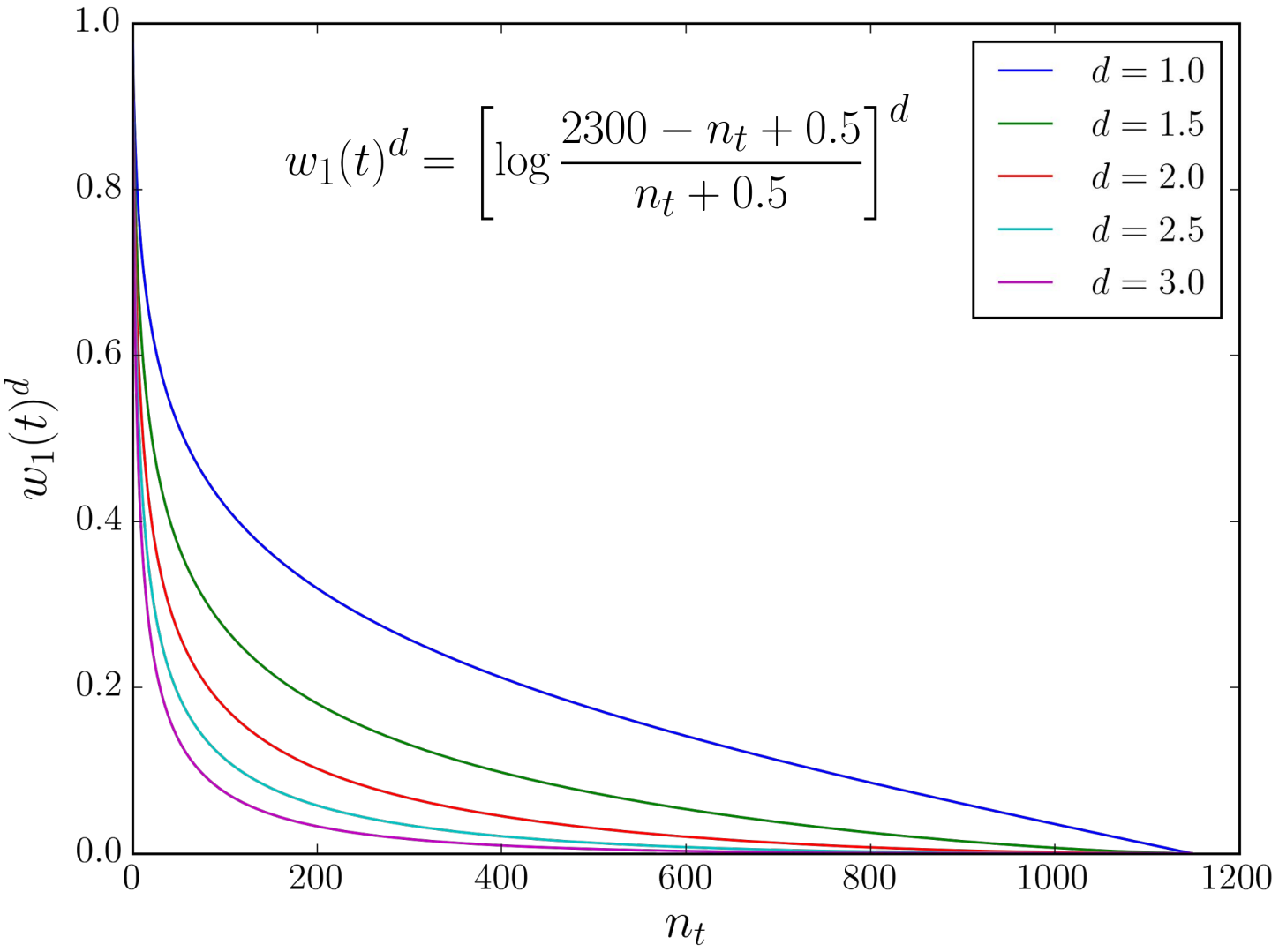
Term weighting

$$w_e(t) = \frac{(k_1+1)tf(t)}{tf(t)+k_1(1-b+b\frac{\text{docl}}{\text{avel}})}\frac{(k_3+1)qf(t)}{qf(t)+k_3}w_1(t)^d$$

IDF exponent

$$w_1(t) = \log\frac{N-n_t+0.5}{n_t+0.5}$$

Retrieval models



Parameter optimisation

- Optimal settings might differ for passages and documents.
- We tune parameters for improving **both** passage and document retrieval jointly.
- 13 parameters to be tuned for **DIS-QE**:
  - BM25: *b, k1, k3, d*
  - QE: R, T
  - DIS: ƛ

  } x 2

- M. Taylor et. al. "*Optimisation methods for ranking functions with multiple parameters*". In Proc. of CIKM'06, pages 585-593, 2006.

Experiments and results

- Parameters trained on SQD-1 queries.
- ASR = REF-WORD-MATCH
- Bold, *, and ^ mark significant differences w.r.t BM25, BM25-QE, and DSI-QE respectively @ 0.95 conf. level.

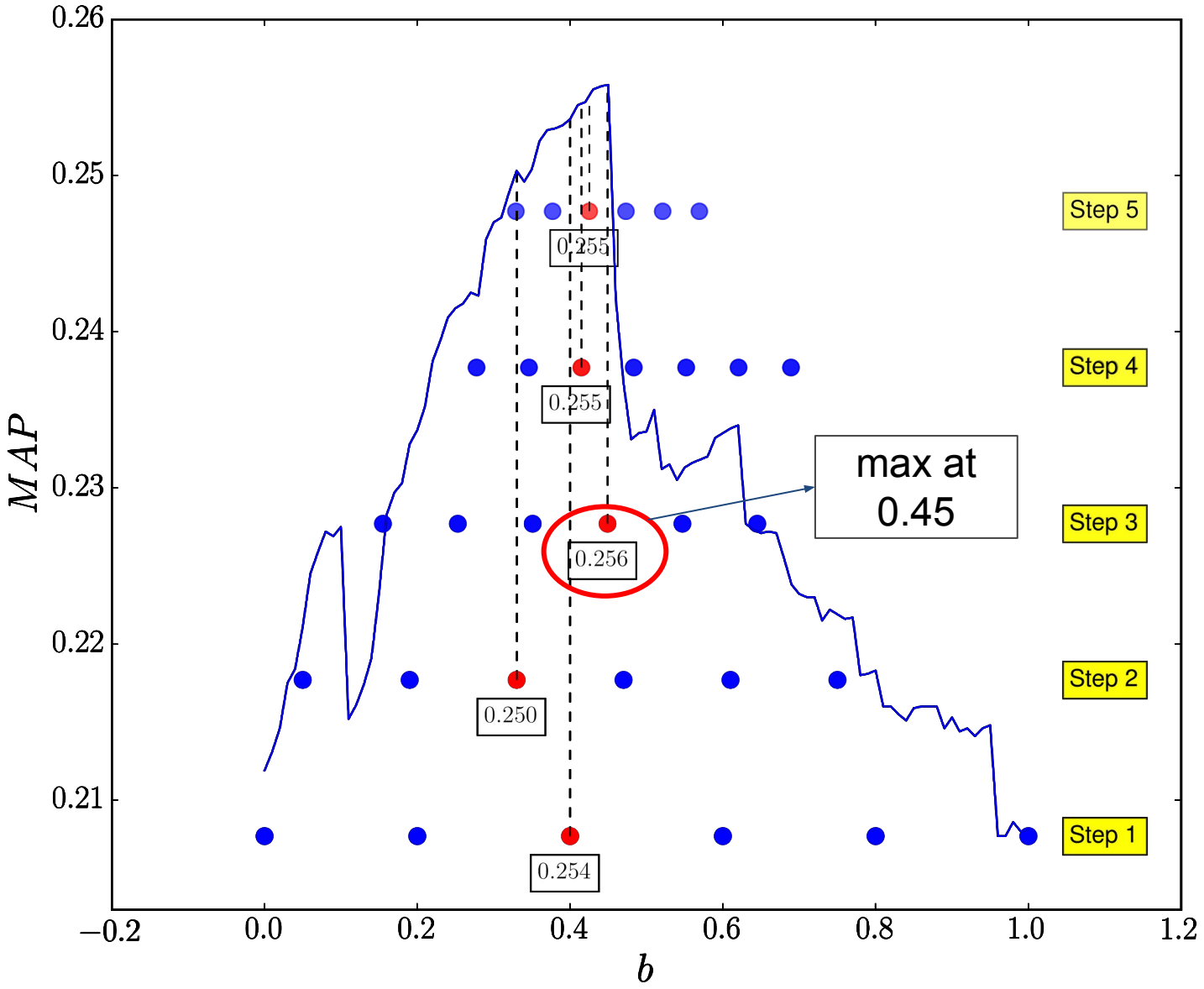
Query	Doc.	Model	MAP
MAN	MAN	BM25	.278
		BM25-QE	.293
		DSI	**.343***
		DSI-QE	**.342***
MAN	ASR	BM25	.212
		BM25-QE	.217
		DSI	**.279*^**
		DSI-QE	.238
ASR	ASR	BM25	.188
		BM25-QE	.183
		DSI	**.250***
		DSI-QE	**.240***

Conclusions

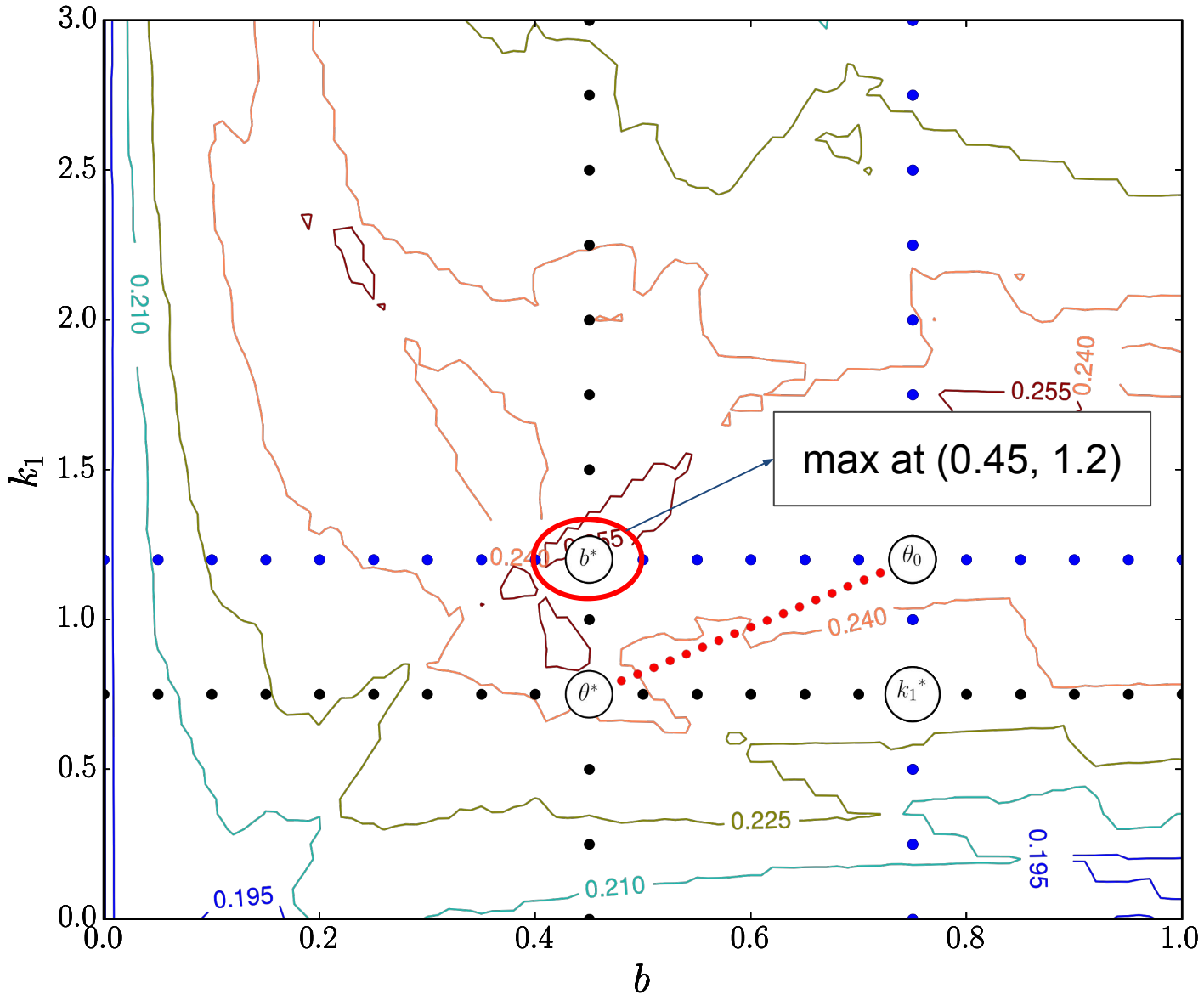
- Results reinforce that passage retrieval effectiveness can be improved if considering document-level evidence.
- Score combination seems to be an effective technique to achieve this.
- PRF and QE were not found to provide significant gains in retrieval effectiveness on this task.

Questions?

Parameter optimisation



Parameter optimisation



Parameter optimisation

Query	Doc	Level	b	k1	k3	d	R	T	ƛ
MAN	MAN	doc.	0.00	4.30	1.27	1.02	4	7	0.70
		pass.	0.38	1.84	254.82	1.03	4	30	
MAN	ASR	doc.	0.02	3.31	1.51	1.00	5	26	0.63
		pass.	0.38	2.52	948.95	1.05	2	5	
ASR	ASR	doc.	0.38	3.48	4.89	1.03	3	28	0.63
		pass.	0.47	2.10	225.00	1.11	2	12	