Information Science
Universität Hildesheim
mandl@uni-hildesheim.de

CLEF 2006
-
Robust Overview

Cross-Language
Evaluation Forum (CLEF)

Why Robust?

The user never sees the perspective of an evaluation (=MAP)

but only the performance on his/her request(s).

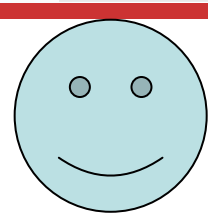
Why Robust?

«The **unhappy customer**, on average, **will tell 27** other people about their experience. With the use of the internet, whether web pages or e-mail, that number can increase to the thousands ...»

«**Dissatisfied customers tell** an average of **ten** other people about their bad experience. Twelve percent tell up to twenty people.»

→ ***Bad news travels fast.***

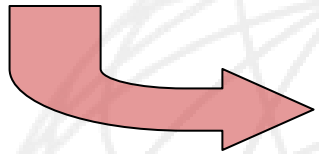
Why Robust?



On the other hand, satisfied customers will tell an average of *five* people about their positive experience.

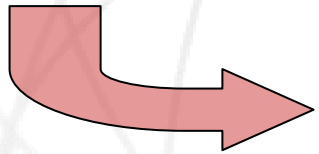
→ ***Good news travels somewhat slower***

Your system should produce less bad news!



improve on *hard* topics

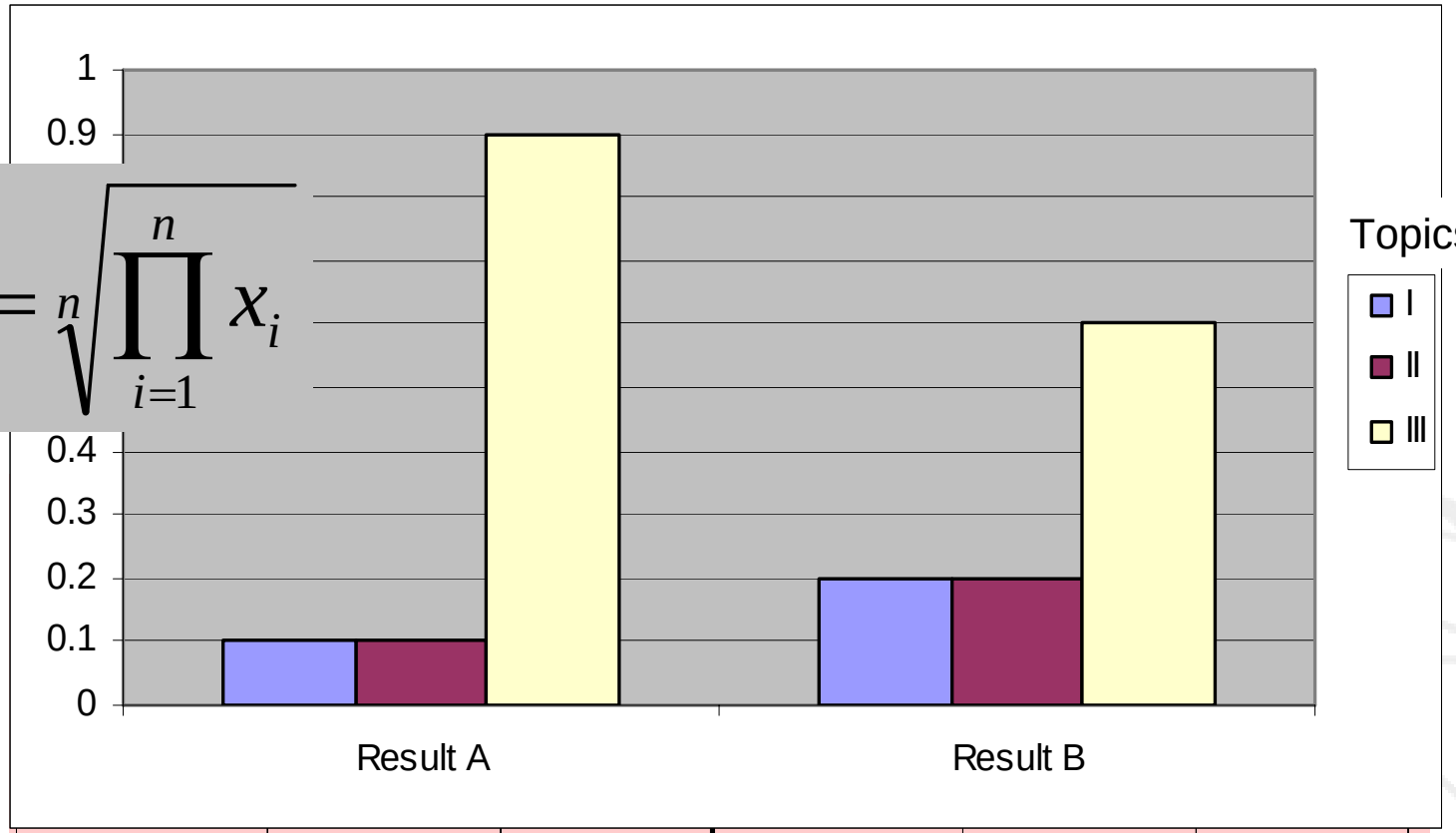
Don't worry too much about the good news



best topics

Which system is better?

$$geoAve = \sqrt[n]{\prod_{i=1}^{n} x_i}$$



GeoAve	A	0.21	GeoAve	B	**0.29**
MAP	A	**0.37**	MAP	B	0.33

Purpose: Robustness in CLIR

- Robustness in multilingual retrieval
  - Stable performance over all topics instead of high average performance (like at TREC, just for other languages)
  - Stable performance over all topics for multi-lingual retrieval
  - Stable performance over different languages (so far at CLEF ?)

Ultimate Goal

- „More work needs to be done on customizing methods for each topic“ (Harman 2005)

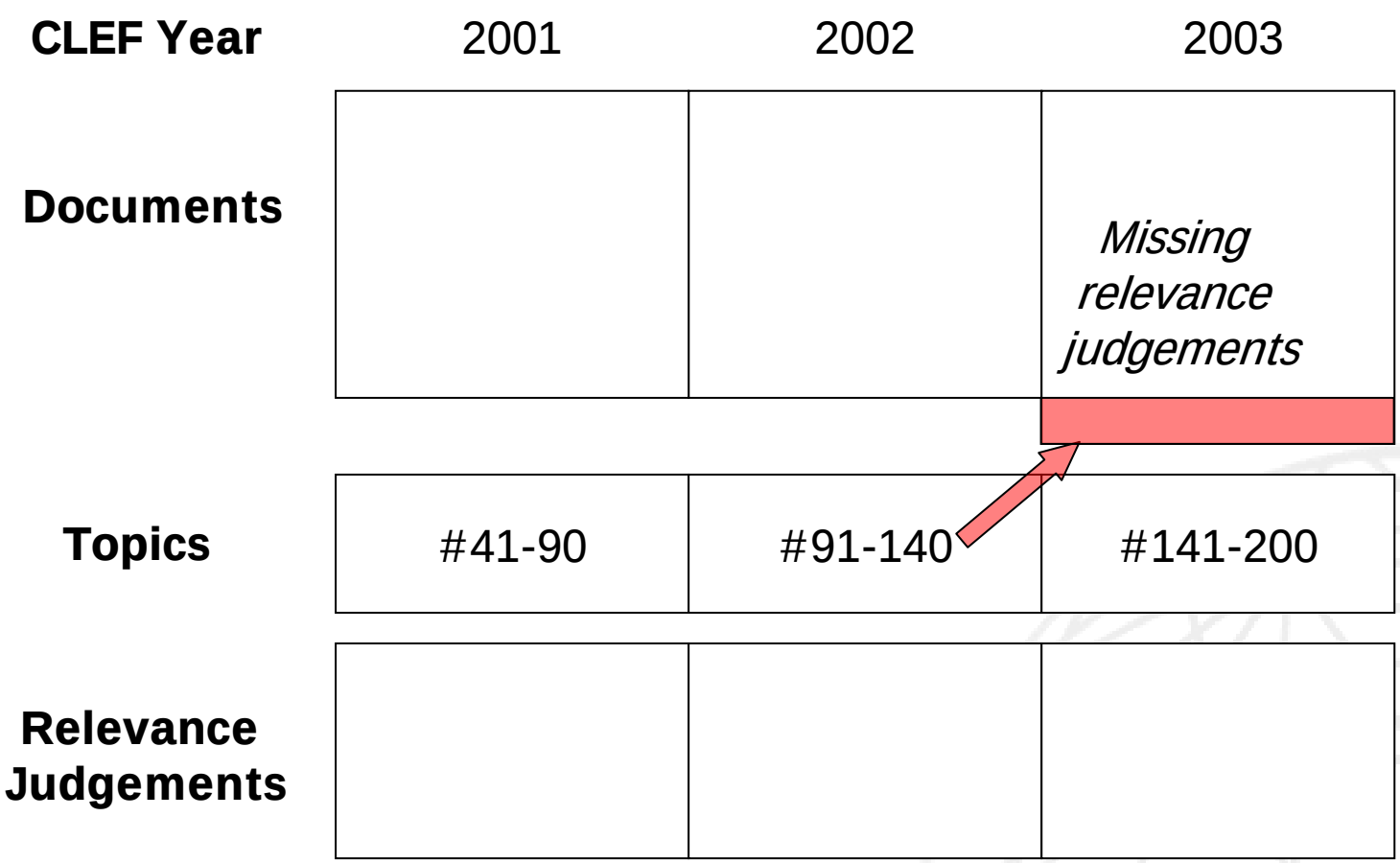
- Hard, but some work ist done
  - RIA Worshop
  - SIGIR Workshop on Topic Difficulty
  - TREC Robust Track (until 2005)
  - ...

Data Collection

- Six languages
- 1.35 million documents
- 3.6 gigabytes of text

Language	Collection
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94
Italian	La Stampa 94, AGZ (SDA) 94/95
Dutch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
German	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanish	EFE 94/95

Data Collections



Test and Training Set

- Arbitrary split of the 160 topics
  - 60 training topics
  - 100 test topics

Identification of Difficult Topics

- Finding a set of difficult topics
  ... was difficult
  - Before the campaign, **no** topics were found which were difficult for more than one language or task at CLEF 2001, CLEF 2002 and CLEF 2003
  - Several definitions of difficulty were tested
- Just as at Robust Track @ TREC
  - Topics are not difficult by themselves
  - but only in interaction with a collection

Task Design for Robustness

- Sub-Tasks
  - Mono-lingual
    - Only for the six document languages as topic language
  - Bi-lingual
    - it->es
    - fr->nl
    - en->de
  - Multi-lingual

Sub-Tasks

- Submission of Training data was encouraged
  - Further analysis of topic difficulty

- Overall, systems did better on test topics

Participants

- U. Coruna & U. Sunderland (Spain & UK)
- U. Jaen (Spain)
- DAEDALUS & Madrid Univs. (Spain)
- U. Salamanca – REINA (Spain)
- Hummingbird Core Tech. (Canada)
- U. Neuchatel (Switzerland)
- Dublin City U. – Computing (Ireland)
- U. Hildesheim – Inf. Sci. (Germany)

More than 100 Runs ...

Task	Language	nr test runs	nr training runs	nr groups
mono	en	13	7	6
	fr	18	10	7
	nl	7	3	3
	de	7	3	3
	es	11	5	5
	it	11	5	5
bi	it->es	8	2	3
	fr->nl	4	0	1
	en->de	5	1	2
multi	multi	10	3	4

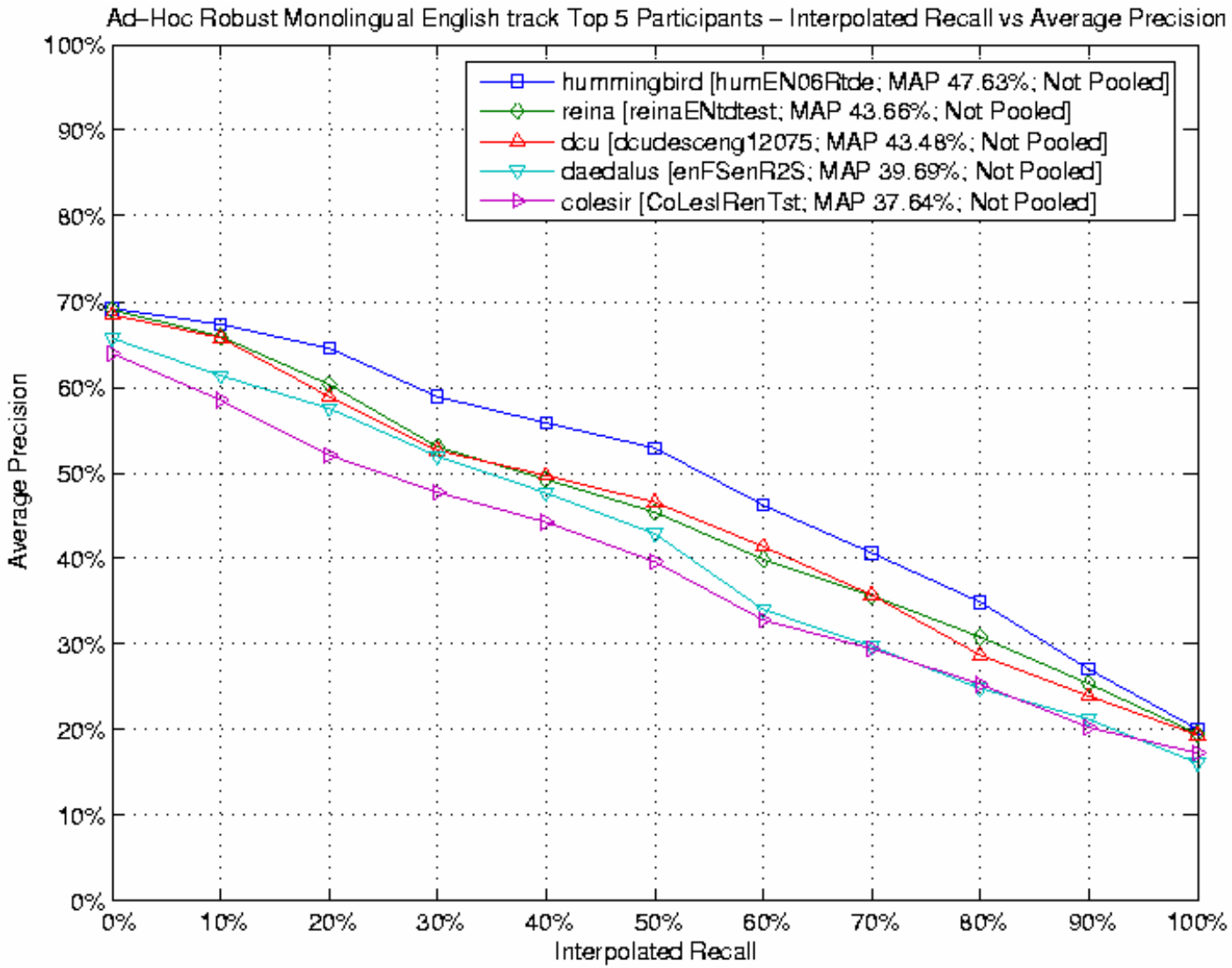
Results – Mono-lingual

Track	Participant Rank					
	1st	2nd	3rd	4th	5th	Diff.
Dutch	hummingbird	daedalus	colesir			1st vs 3rd
MAP	51.06%	42.39%	41.60%			22.74%
GMAP	25.76%	17.57%	16.40%			57.13%
Run	humNL06Rtd	nlFSnlR2S	CoLesIRnlTst			
English	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	47.63%	43.66%	43.48%	39.69%	37.64%	26.54%
GMAP	11.69%	10.53%	10.11%	8.93%	8.41%	39.00%
Run	humEN06Rtd	reinaENtdtest	dcudesceng1	enFSenR2S	CoLesIRenTst	
French	unine	hummingbird	reina	dcu	colesir	1st vs 5th
MAP	47.57%	45.43%	44.58%	41.08%	39.51%	20.40%
GMAP	15.02%	14.90%	14.32%	12.00%	11.91%	26.11%
Run	UniNEfrr1	humFR06Rtd	reinaFRtdtest	dcudescfr120	CoLesIRfrTst	

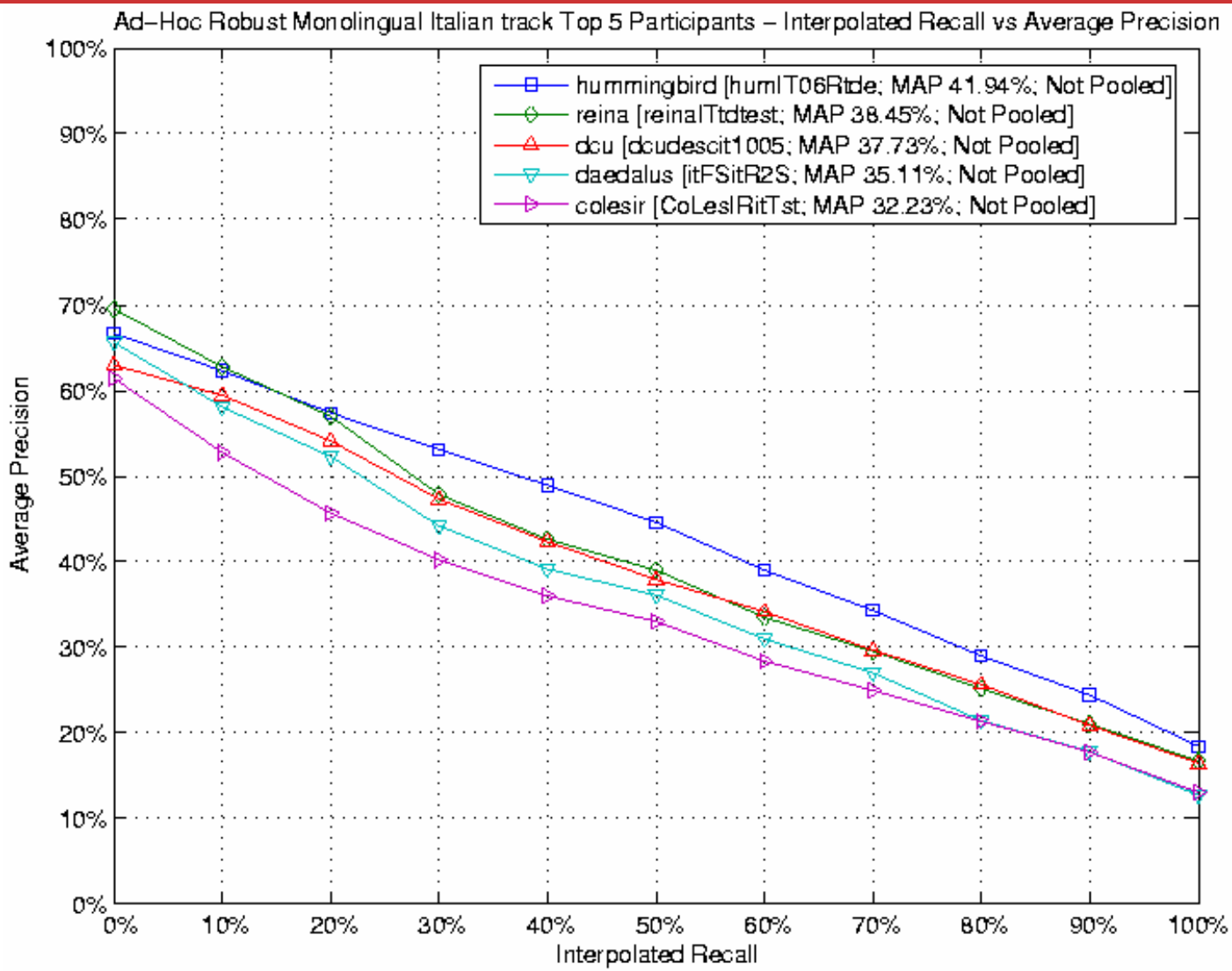
Results – Mono-lingual

Track	Participant Rank					
	1st	2nd	3rd	4th	5th	Diff.
German	hummingbird	colesir	daedalus			1st vs 3rd
MAP	48.30%	37.21%	34.06%			41.81%
GMAP	22.53%	14.80%	10.61%			112.35%
Run	humDE06Rtd	CoLesIRdeTs	deFSdeR2S			
Italian	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	41.94%	38.45%	37.73%	35.11%	[illegible]	30.13%
GMAP	11.47%	10.55%	9.19%	10.50%	[illegible]	39.37%
Run	humIT06Rtde	reinaITtdtest	dcu[illegible]	[illegible]	CoLesIRitTst	
Spanish	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	45.66%	44.01%	42.14%	40.40%	40.17%	13.67%
GMAP	23.61%	22.65%	21.32%	19.64%	18.84%	25.32%
Run	humES06Rtd	reinaEStdtest	dcudescsp12	esFSesR2S	CoLesIResTst	

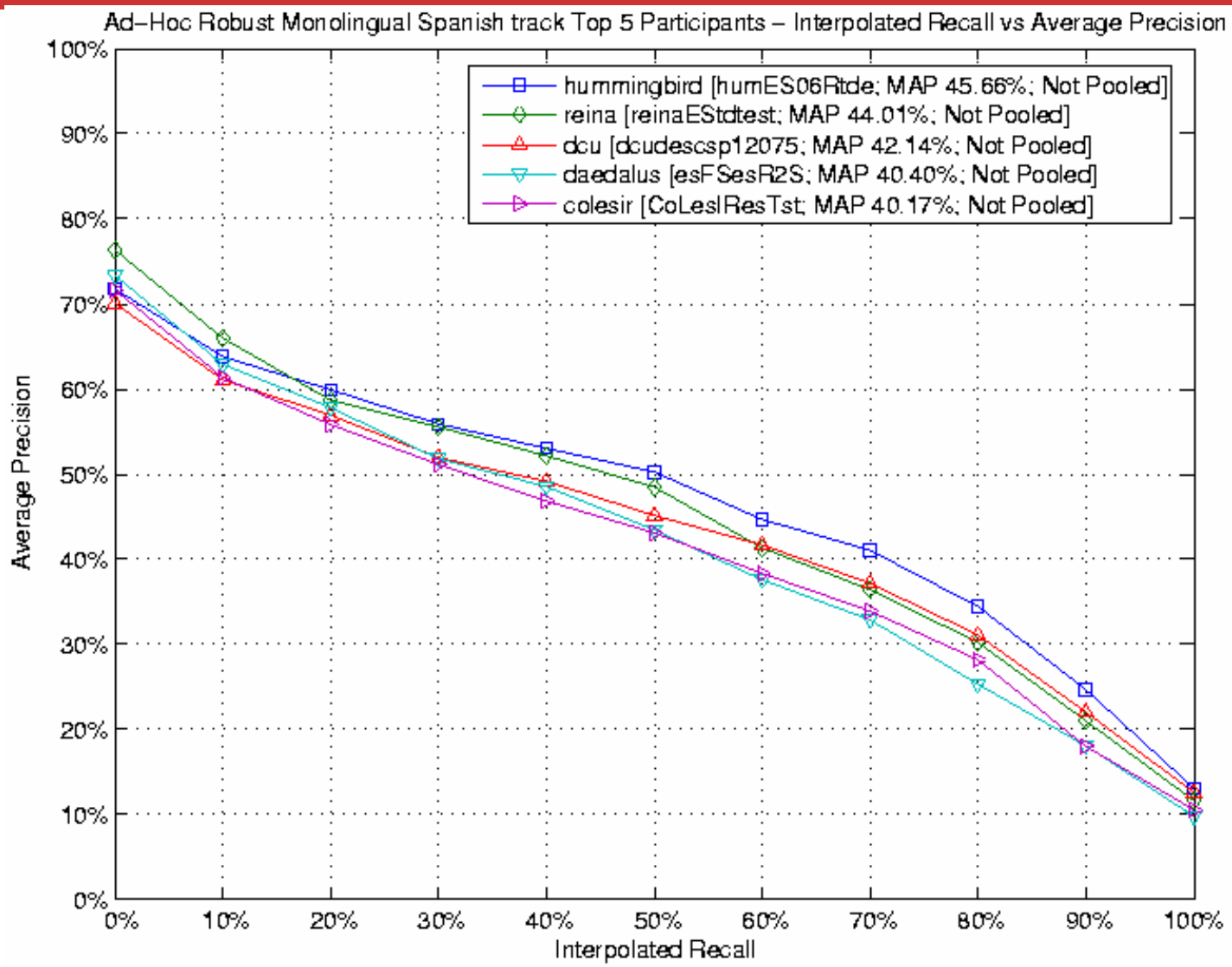
Results – Mono-lingual English



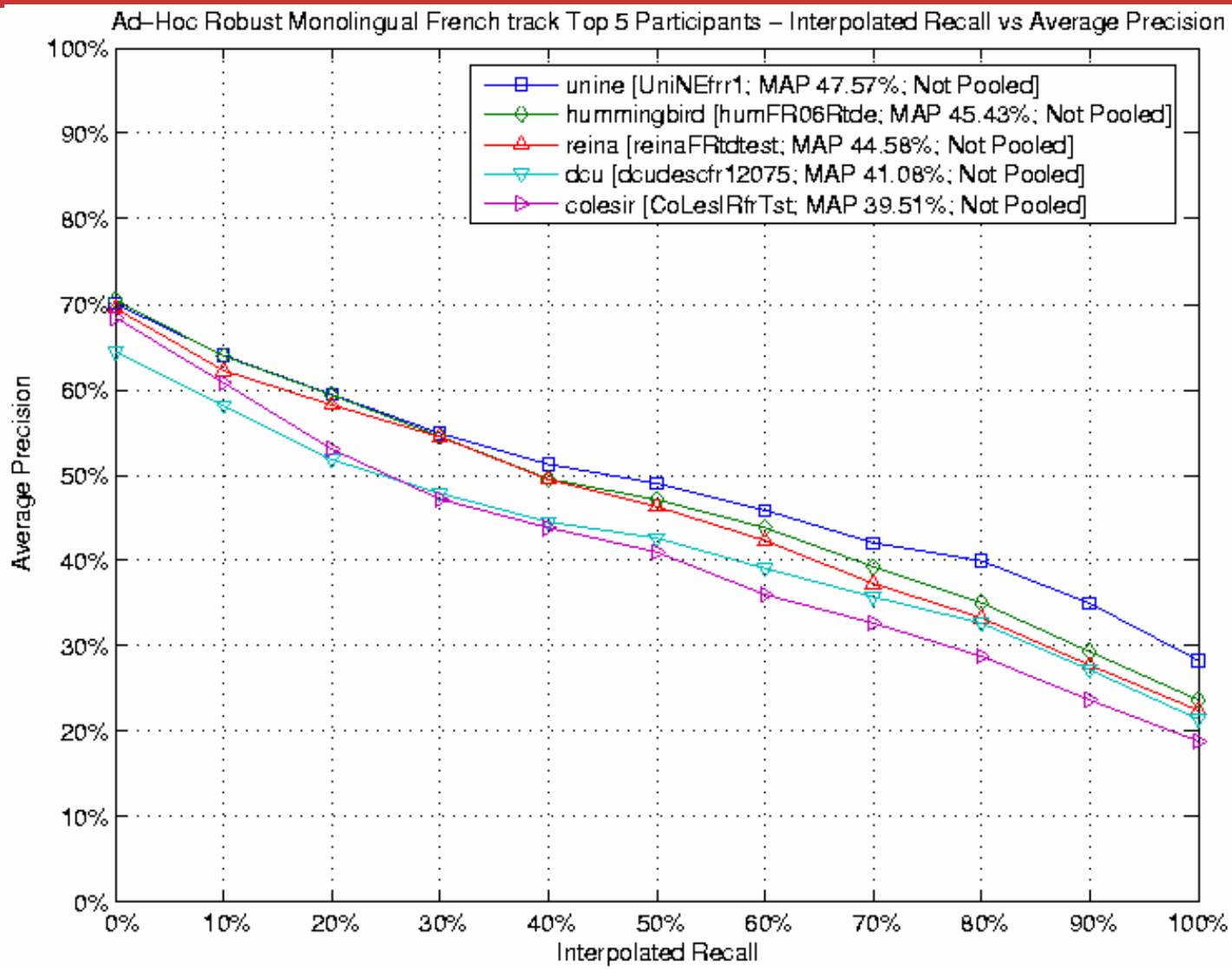
Results – Mono-lingual Italian



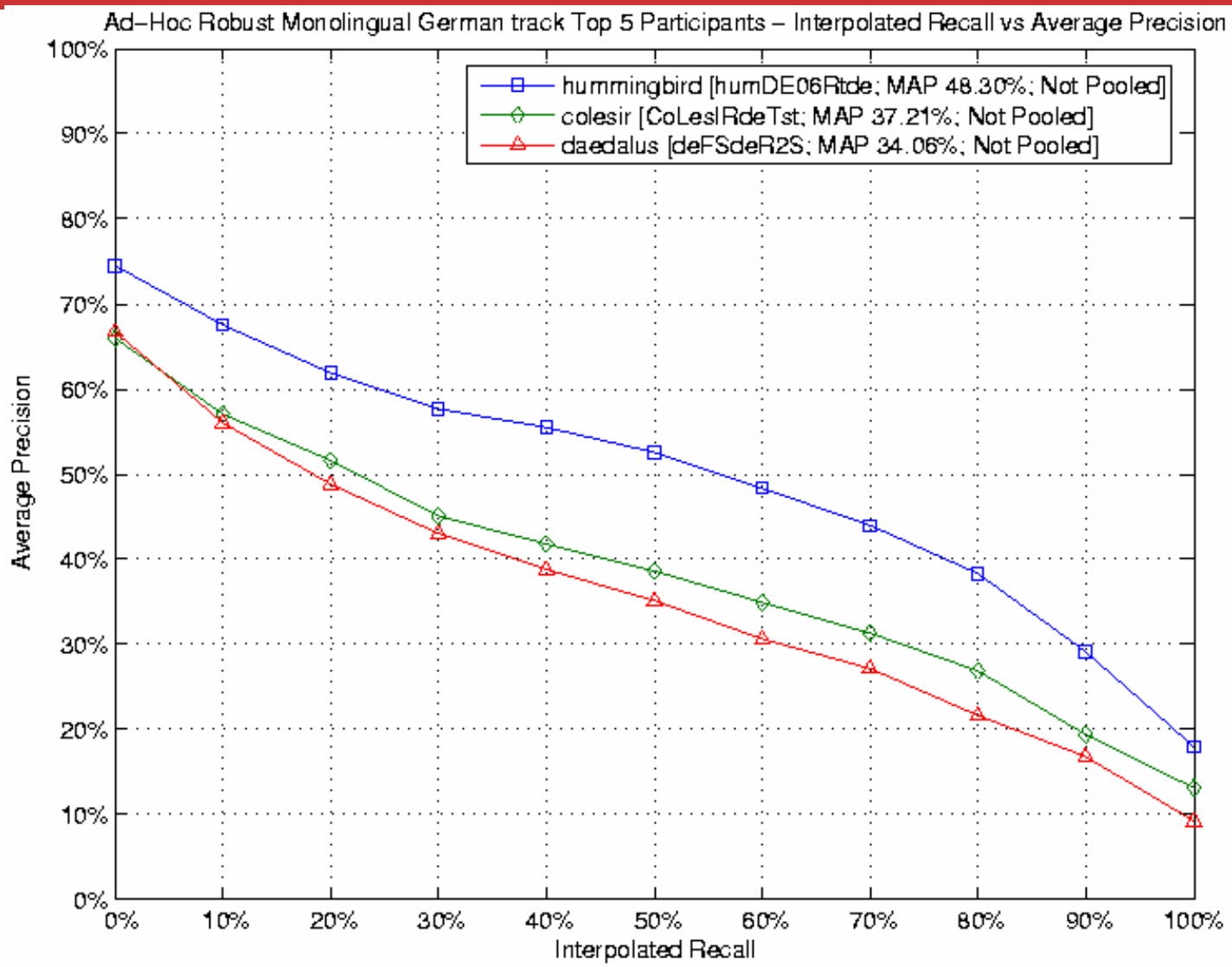
Results – Mono-lingual Spanish



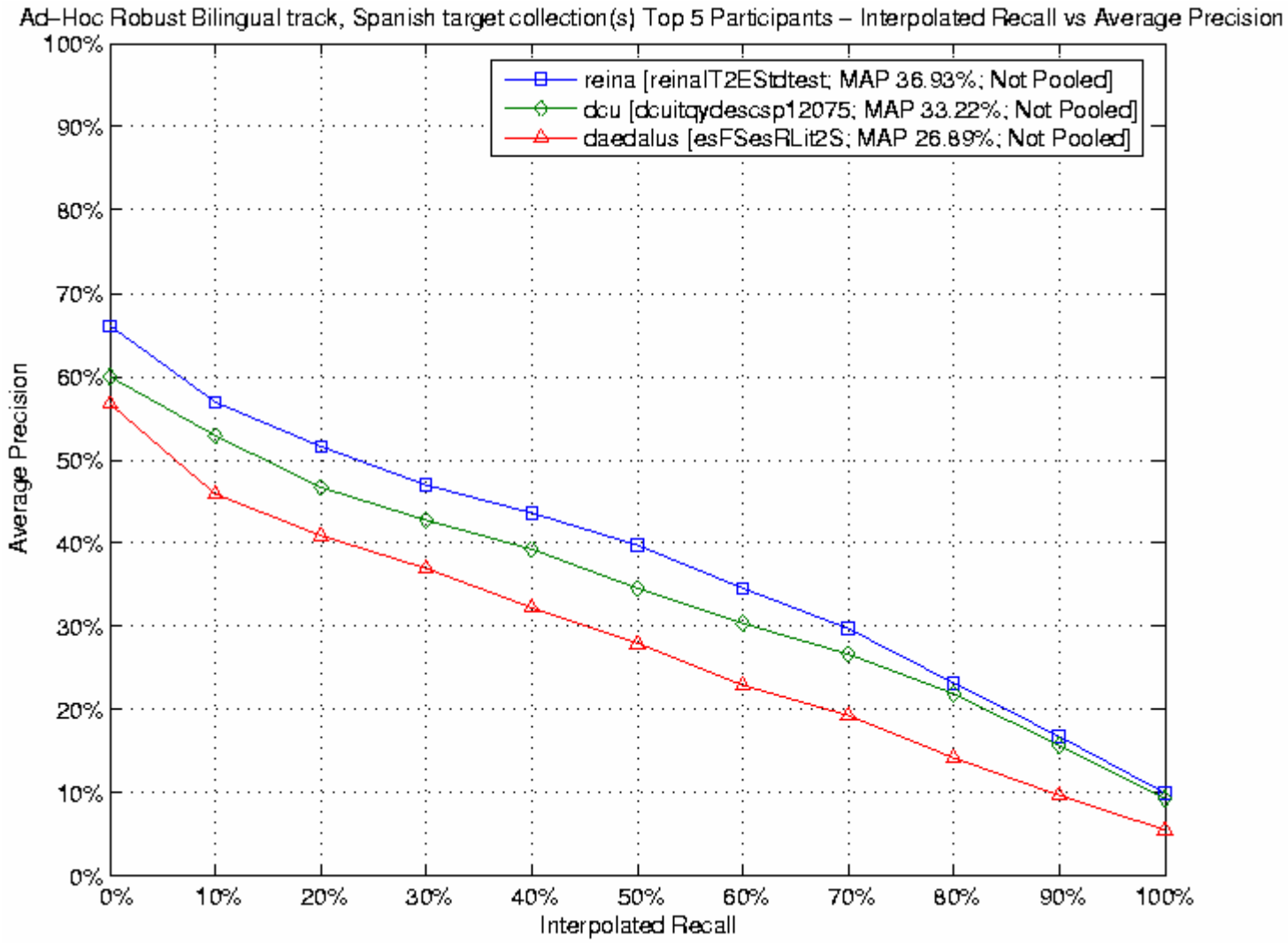
Results – Mono-lingual French



Results – Mono-lingual German



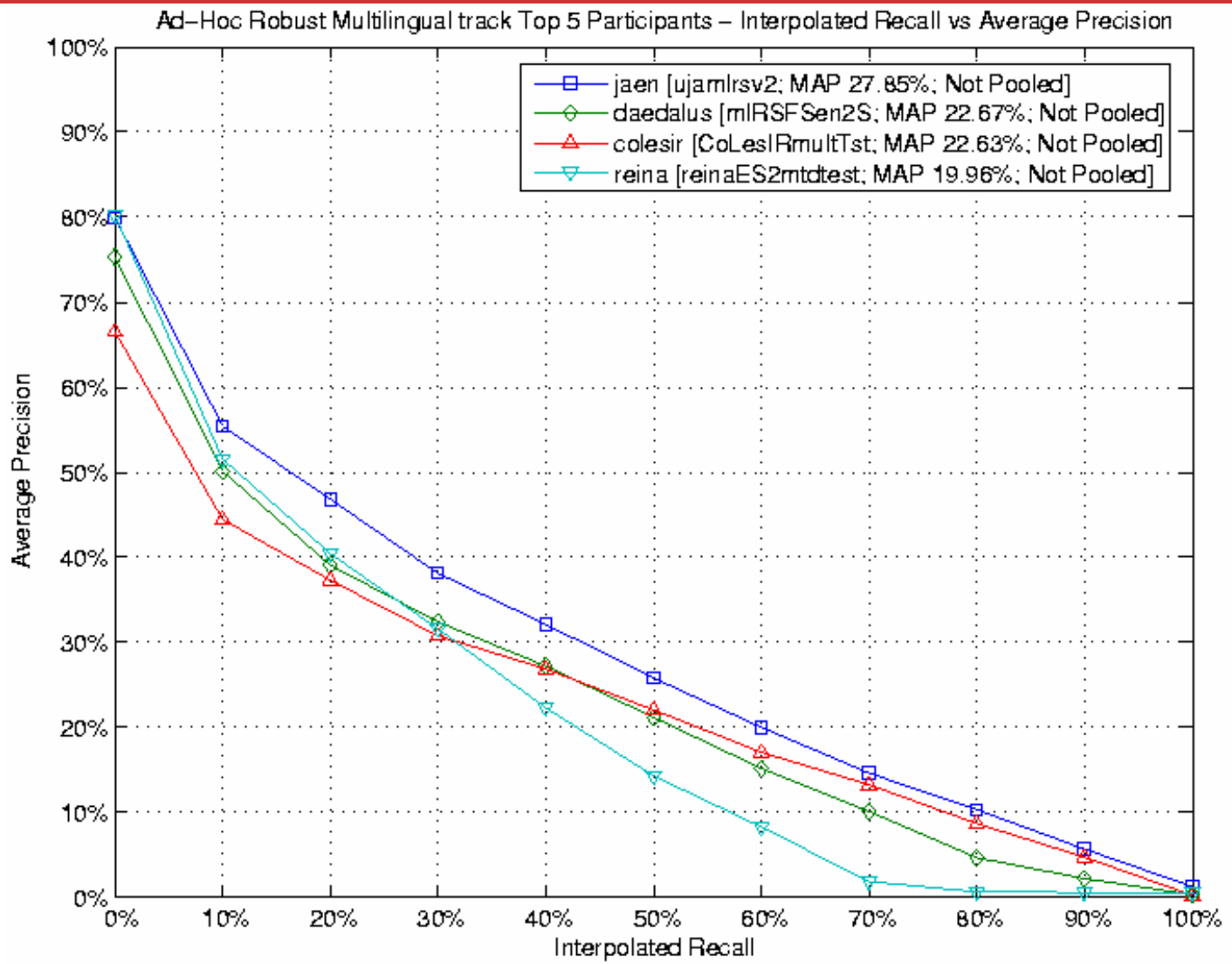
Results – Bi-lingual



Results – Multi-lingual

Track	Participant Rank			
	1st	2nd	3rd	4th
Multilingual	jaen	daedalus	colesir	reina
MAP	27.85%	22.67%	22.63%	19.96%
GMAP	15.69%	11.04%	11.24%	13.25%
Run	ujamlrsv2	mlRSFSen2S	CoLesIRmultTst	reinaES2mtdtest

Results – Multi-lingual



Topic Analysis

- Example: topic 64
  - Easiest topic for mono and bi German
  - Hardest topic for mono Italian
- Example: topic 144
  - Easiest topic for four bi Dutch
  - hardest for mono and bi German
- Example: topic 146
  - Among three hardest topics for four sub-tasks
  - mid-range for all other sub-tasks

Approaches

- SINAI expanded with terms gathered from a web search engine [Martinez-Santiago et al. 2006]
- REINA used a heuristic for determining hard topics during training. Different expansion techniques were applied [Zazo et al. 2006]
- Hummingbird experimented with other evaluation measures then used in the track [Tomlinson 2006].
- MIRACLE tried to find a fusion scheme which is good for the robust measure [Goni-Menoyo et al. 2006]

Outlook

- What can we do with the data?
- Have people improved in comparison to CLEF 2001 through 2003?
- Are low MAP scores a good indicator of topic difficulty?

Acknowledgements

- Giorgio di Nunzio & Nicola Ferro (U Padua)
- Robust Committee
  - Donna Harman (NIST)
  - Carol Peters (ISTI-CNR)
  - Jacques Savoy (U Neuchâtel)
  - Gareth Jones (Dublin City U)
- Ellen Voorhees (NIST)
- Participants

Thanks for your Attention

I am looking forward to the Discussion

Please come to the Breakout Session