# NIST OpenHaRT'13 Evaluation: Overview and Results

Audrey Tong

NIST 2013 Open Handwriting Recognition and Translation (OpenHaRT'13) Evaluation Workshop
Omni Shoreham Hotel, Washington DC
August 23, 2013

# Outline

- Introduction
- Schedule
- Tasks
- Training conditions
- Test data
- Metrics
- Participation & submissions
- Dry run
- Evaluation results
- Summary

# Introduction: What Is OpenHaRT?

- Stand for Open Handwriting Recognition and Translation evaluation
- Modeled after the evaluation conducted for the DARPA Multilingual Automatic Document Classification Analysis and Translation (MADCAT) Program
- Currently focus on technology that recognizes Arabic script in images and translates it into English
- Would like to extend to other technologies that contribute toward document understanding
- Open to the public to build critical mass to solve the technical challenges
- Had its first evaluation in 2010

# Introduction: Goal

- Support research and help advance the state-of-the-art of document understanding technologies
  - Define a set of tasks
  - Collaborate with data providers to produce data resources to support research on these tasks
  - Manage the evaluations
  - Provide evaluation utilities and infrastructure for researchers to evaluate their techniques
  - Coordinate workshop to discuss findings and guide research directions

# Schedule

| | |
|---|---|
| June 1, 2012 | Evaluation plan posted |
| June 1 – December 31, 2012 | Registration period (training data available few weeks after registration) |
| Feb 14 – 28, 2013 | Dry run evaluation period |
| Apr 9 – May 14, 2013 | Formal evaluation period |
| May 28, 2013 | Preliminary results released |
| June 15, 2013 | System description due |
| August 23, 2013 | Workshop |
| October 23, 2013 | Official results published |

# Tasks

- Document Image Recognition (DIR) [OCR]
  - Given the image and its text line segmentation, recognize the Arabic text in the image
- Document Image Translation (DIT) [OCR+MT]
  - Given the image and its text line segmentation, translate the Arabic text in the image into accurate and fluent English text
- Document Text Translation (DTT) [MT]
  - Given the ground truth Arabic text in the image, translate the Arabic text into accurate and fluent English text

# Training Conditions

- Constrained
  - Only data from the LDC
  - Allow for direct comparison of different algorithmic approaches
- Unconstrained
  - All publicly available data allowed for development except data from test epoch
  - Encourage creativity with as few restrictions as possible

# Test Sets

| Data Sets | | No. of Documents | No. of Segments | No. of Tokens |
|---|---|---|---|---|
| Dry run | Newswire | 267 | 1143 | 33045 |
| | Web | 267 | 1605 | 31863 |
| | Total | 534 | 2748 | 64908 |
| Formal | Newswire | 330 | 1338 | 38466 |
| | Web | 303 | 1806 | 36087 |
| | Total | 633 | 3144 | 74553 |

# Test Data – Sources

- Test data came from Arabic news-related passages obtained from a variety of sources published in June 1-30, 2008
  - newswire publications
  - web postings (blogs & newsgroup discussion forums)
- News-related genres were used to leverage existing data resources (reference translation already exists)

# Test Data – Document Image Creation

- Native Arabs proficient in Arabic were recruited to produce handwritten copies of these passages
- Each passage was copied by three scribes under various conditions

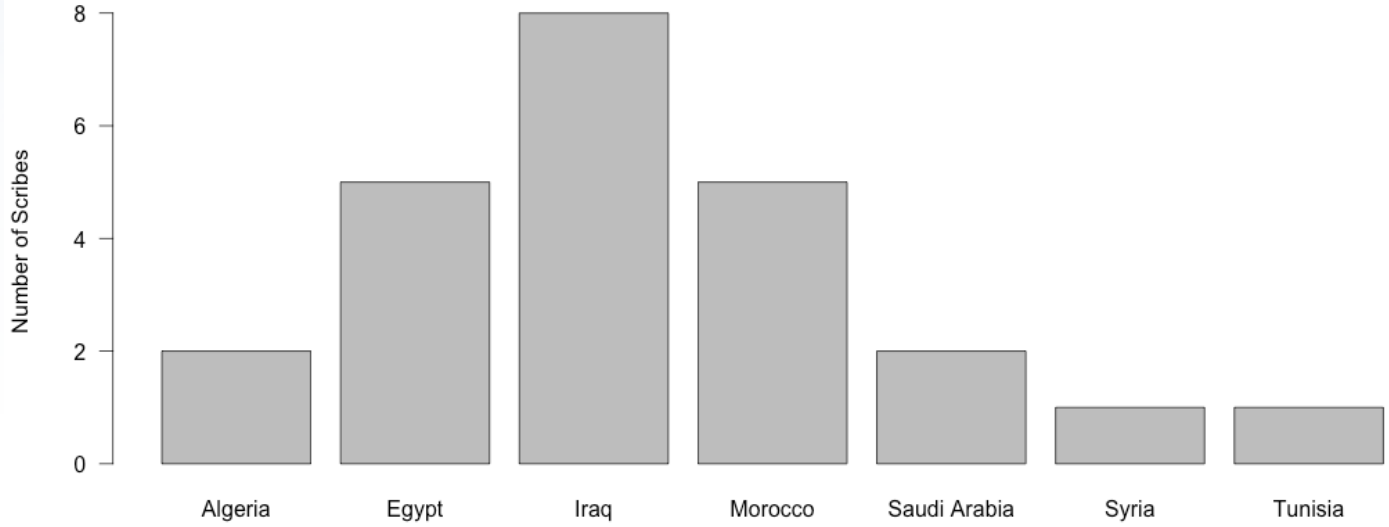
| Writing Instrument | Writing Surface | Writing Speed |
|---|---|---|
| 90% ballpoint pen | 75% unlined paper | 90% normal |
| 10% pencil | 25% lined paper | 5% fast |
| | | 5% careful |

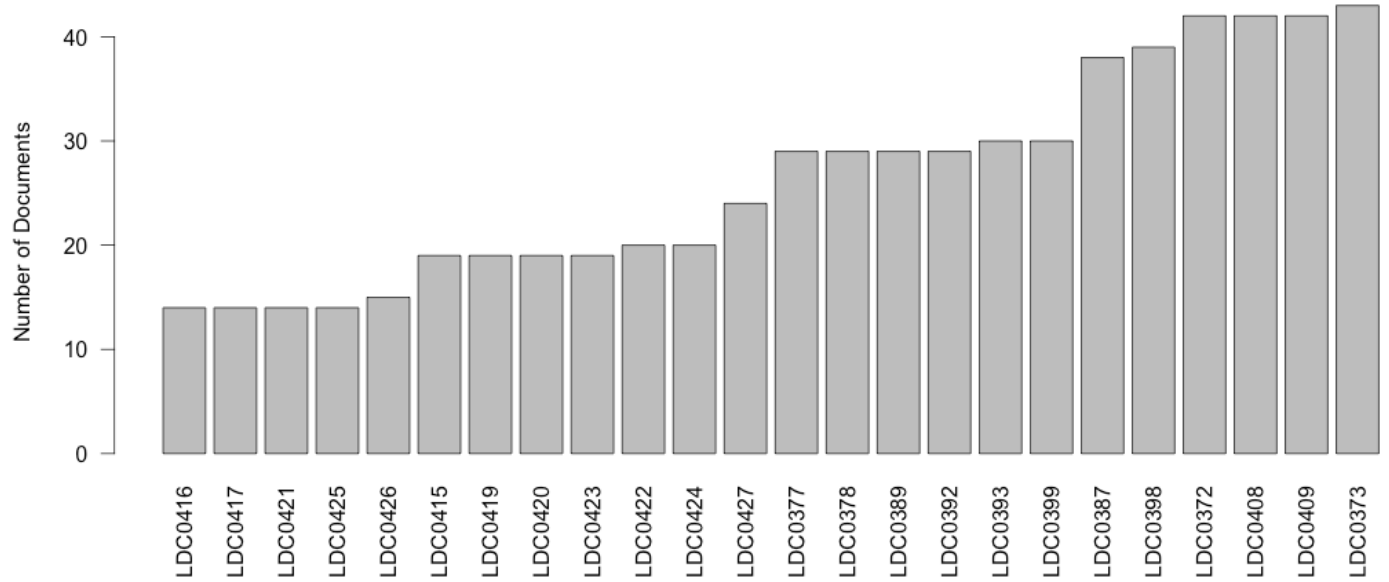
- Documents were scanned at 600 dpi

# Test Data – Size

| Genre | No. of Passages | No. of Pages | No. of Documents | No. of Segments | No. of Tokens/ Words |
|---|---|---|---|---|---|
| Newswire | 66 | 110 | 330 | 1338 | 38466 |
| Web | 57 | 101 | 303 | 1806 | 36087 |
| Total | 123 | 211 | 633 | 3144 | 74553 |

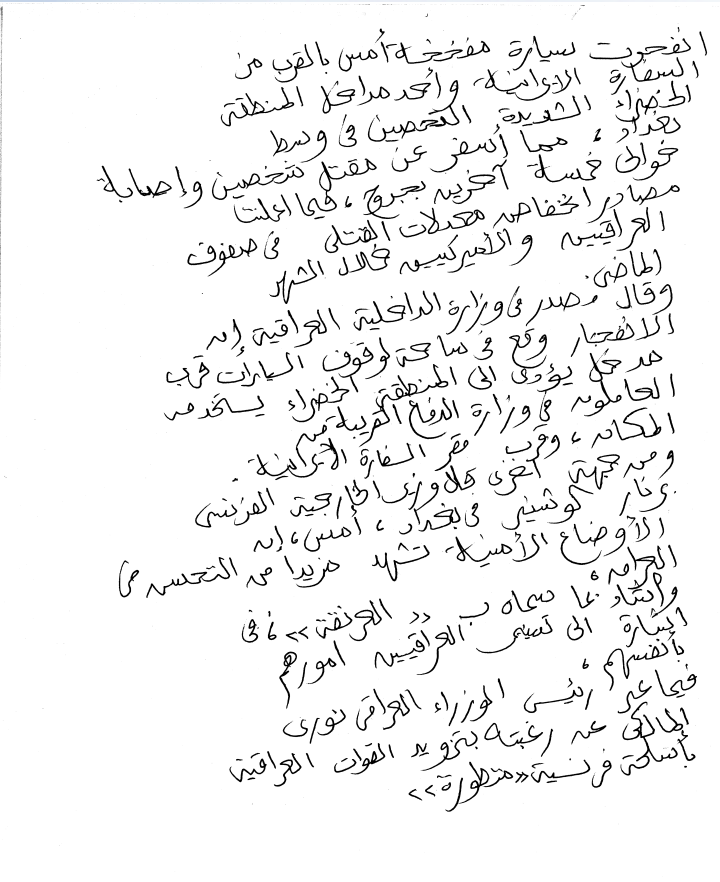
# Test Data – Scribes



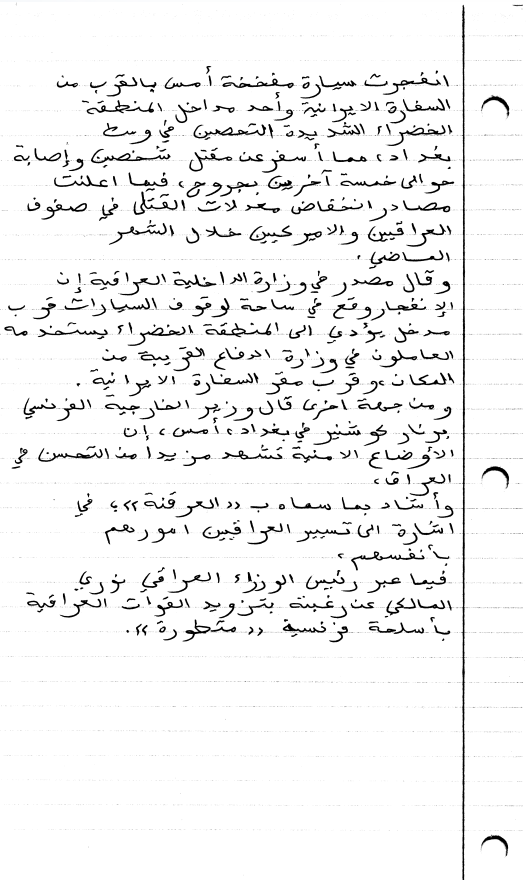




- 24 unique native Arabs from various countries
- Not all scribes wrote the same number of documents

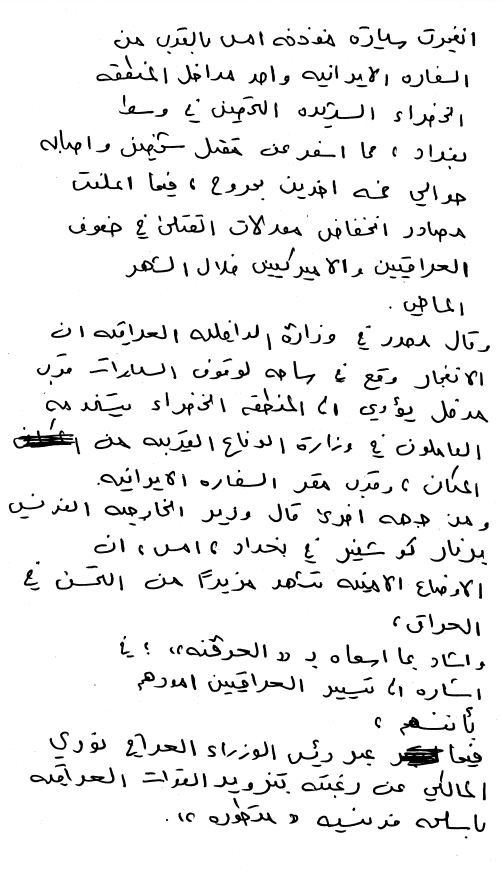
# Test Data – Samples



Pen, Unlined, Normal



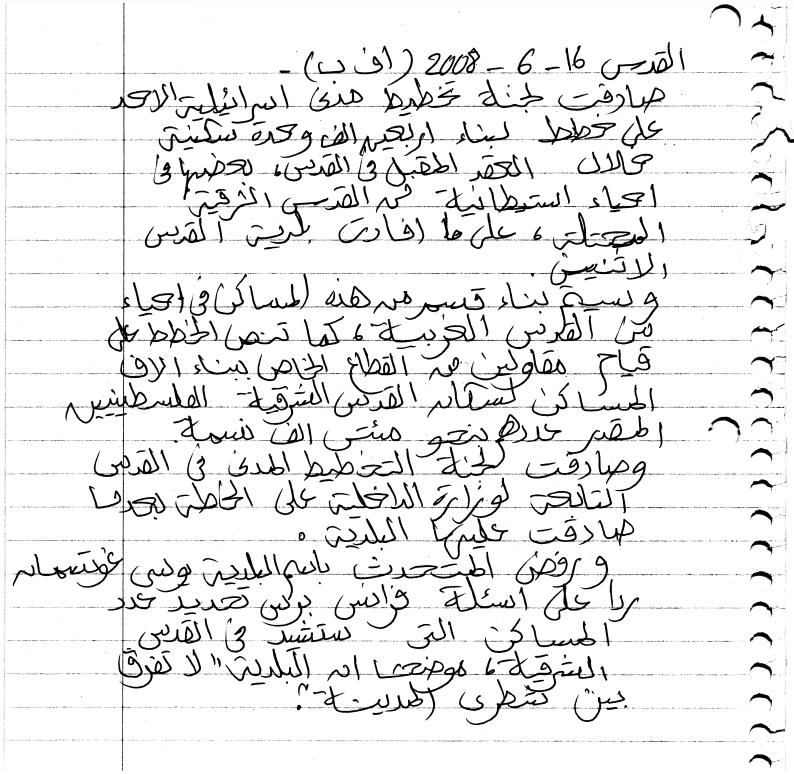
Pen, Lined, Normal



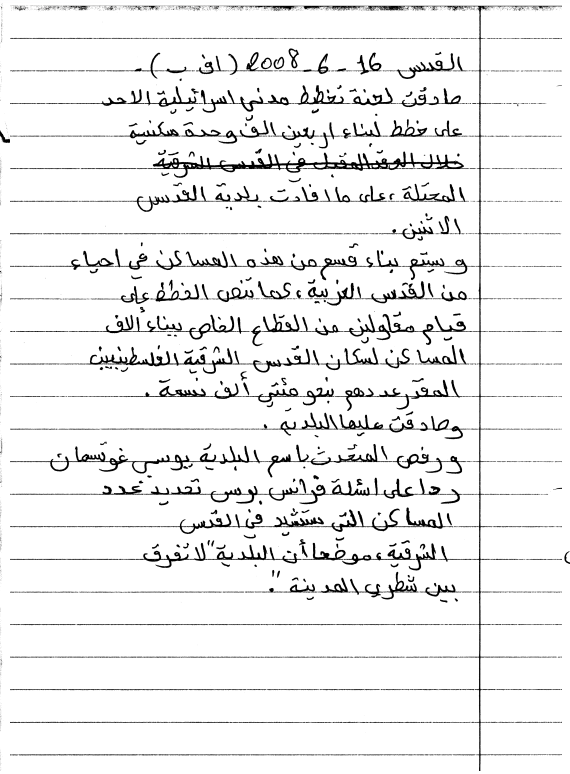
Pen, Unlined, Normal

- Same document different scribes

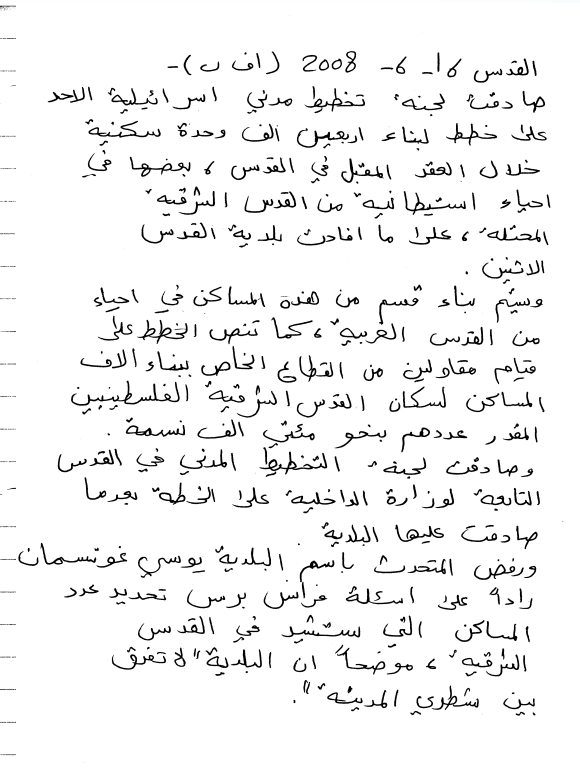
# Test Data – More Samples



Pen, Lined, Fast



Pen, Lined, Normal



Pen, Unlined, Fast

- Same document different scribes

# Metrics

| Task | Metric | Software Version | Author | Description |
|---|---|---|---|---|
| DIR | WER (primary) | sctk-2.4.7 | NIST | Counts of the number of edits required to match the reference transcription |
| DIT DTT | TER (primary) | tercom-0.7.25 | UMD & BBN | Counts of the number of edits required to match the reference translation. Block move is allowed and counted as one edit |
| | BLEU | mteval_v13a | NIST | NIST implementation of IBM's BLEU (n-gram co-occurrence statistics plus brevity penalty) |
| | METEOR | meteor-1.4 | CMU | Unigram matches through a sequence of staging (exact match, Porter stemmer, WordNet stemmer, WordNet synonym) |

# Participation and Submissions

| Team ID | Team Name | DIR | DIT | DTT |
|---|---|---|---|---|
| A2IA | A2iA, France | 2C<br>1U | | |
| CITLAB | Rostock University, Germany | 8C | | |
| LIP6 | University Pierre & Marie Curie, France | | | |
| LITIS | University of Rouen, France | 2C | | |
| MENASRI | Unaffiliated, France | | | |
| RWTH | Aachen University, Germany | 4C<br>1U | | |
| UOB-TPT | University of Balamand, Lebanon &<br>Telecom ParisTech, France | 4C | 1C | 1C<br>1U |
| UPV | Polytechnical University of Valencia, Spain | 2C | 3C | 1C<br>1U |

- Total number of submissions: 32
- No late submission!

orange = newcomer
gray = dropout
C = constrained
U = unconstrained

# Dry Run

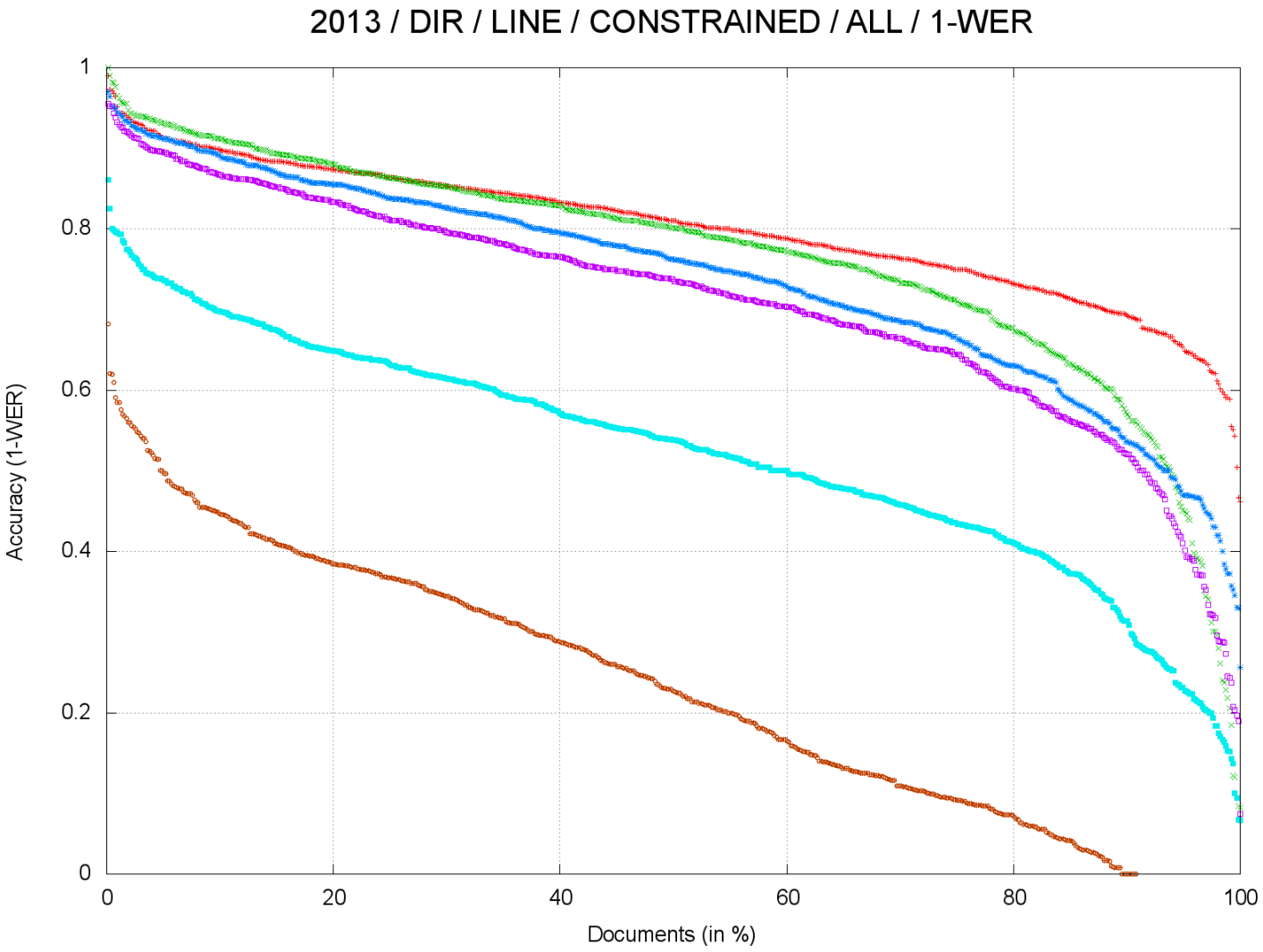
- Teams were required to participate in a dry run evaluation prior to the real evaluation
- Dry run evaluation served as a practice test to iron out any issues prior to the real evaluation
  - Followed the same protocol (as much as possible) to the real evaluation
  - Dry run test data had similar properties as those in the official test set

# Evaluation Results

- Limited to primary systems
- By task & training condition
- By genre
- By metric (translation tasks only)
- Scribe effect
- Comparison against 2010 results

# Results: DIR Constrained Training

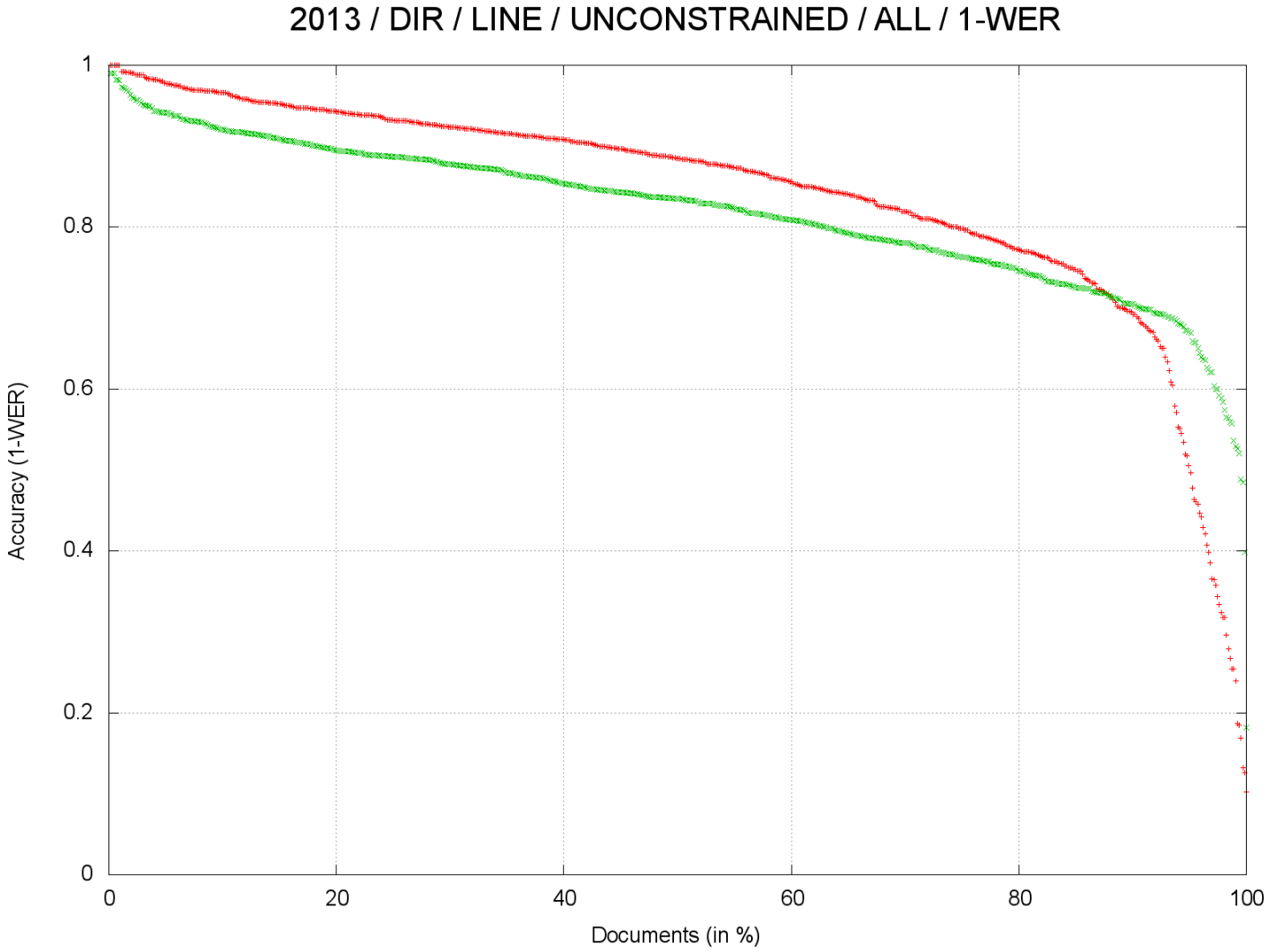
## Document Image Recognition (OCR)




| | Avg Score |
|---|---|
| A2IA | 0.80 |
| RWTH | 0.76 |
| CITLAB | 0.74 |
| UPV | 0.71 |
| UOB-TPT | 0.52 |
| LITIS | 0.22 |

# Results: DIR Unconstrained Training

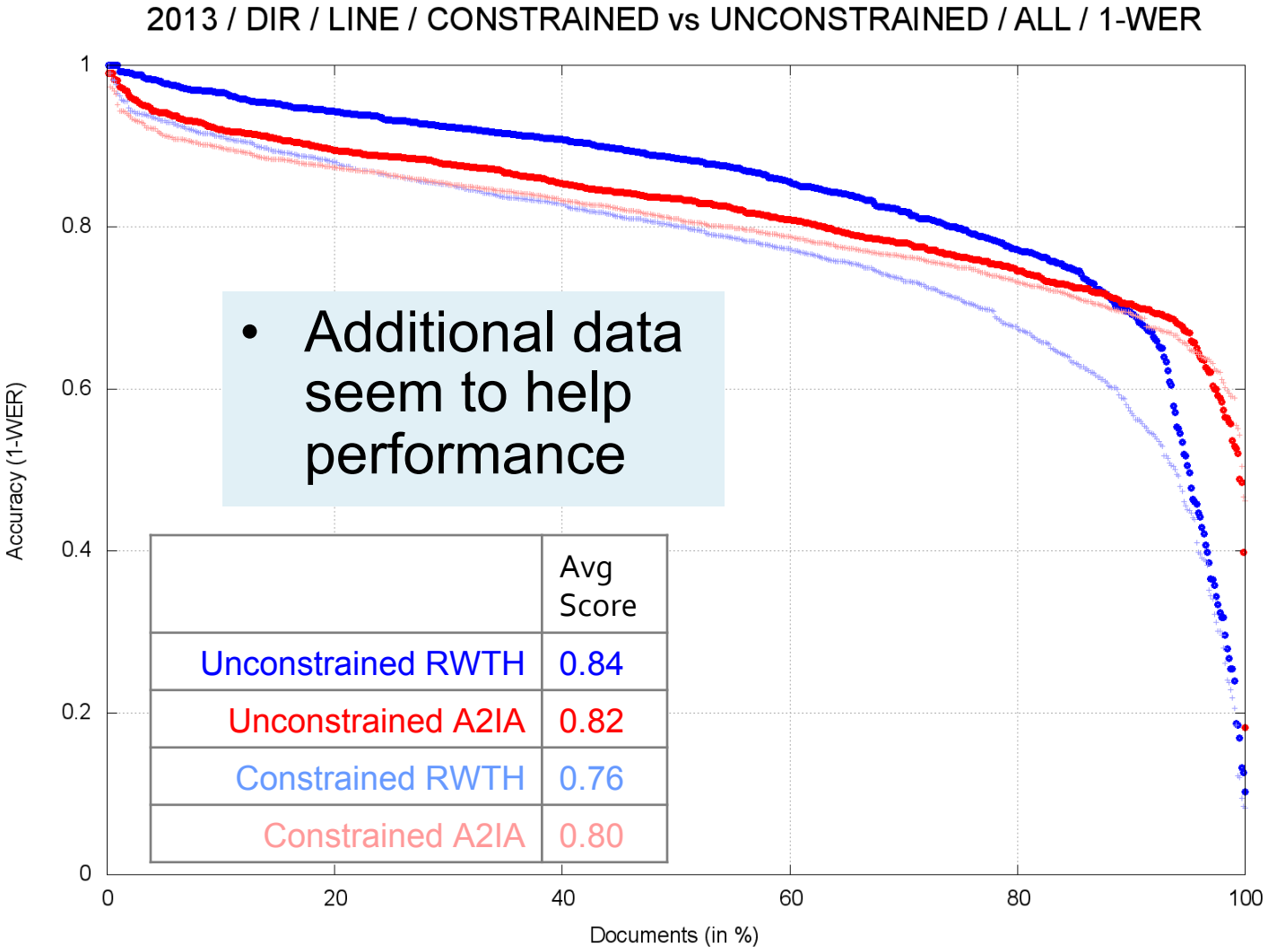
## Document Image Recognition (OCR)




| | Avg Score |
|---|---|
| RWTH | 0.84 |
| A2IA | 0.82 |

# Results: DIR Constrained vs. Unconstrained

## Document Image Recognition (OCR)

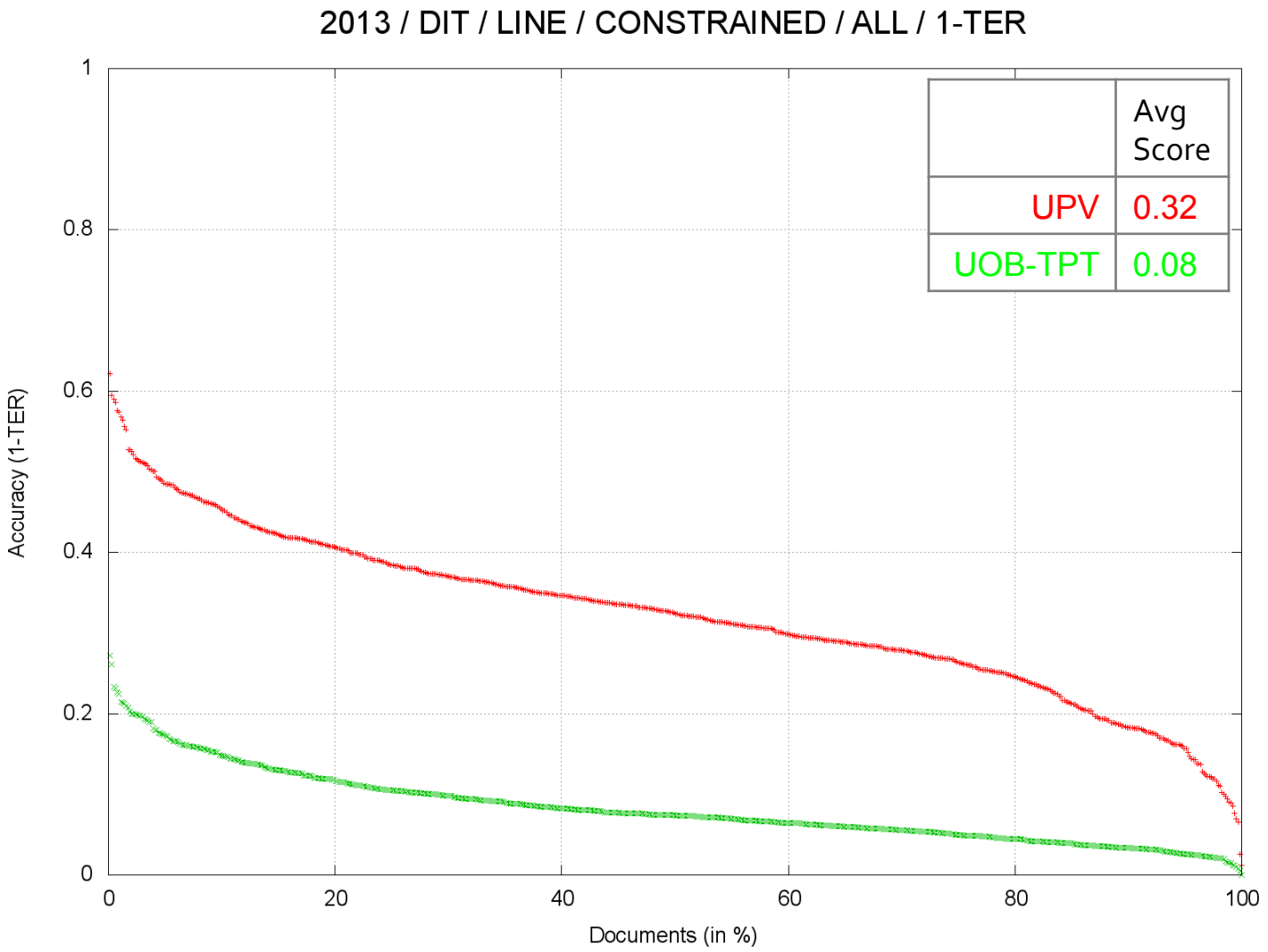



- Additional data seem to help performance

| | Avg Score |
|---|---|
| Unconstrained RWTH | 0.84 |
| Unconstrained A2IA | 0.82 |
| Constrained RWTH | 0.76 |
| Constrained A2IA | 0.80 |

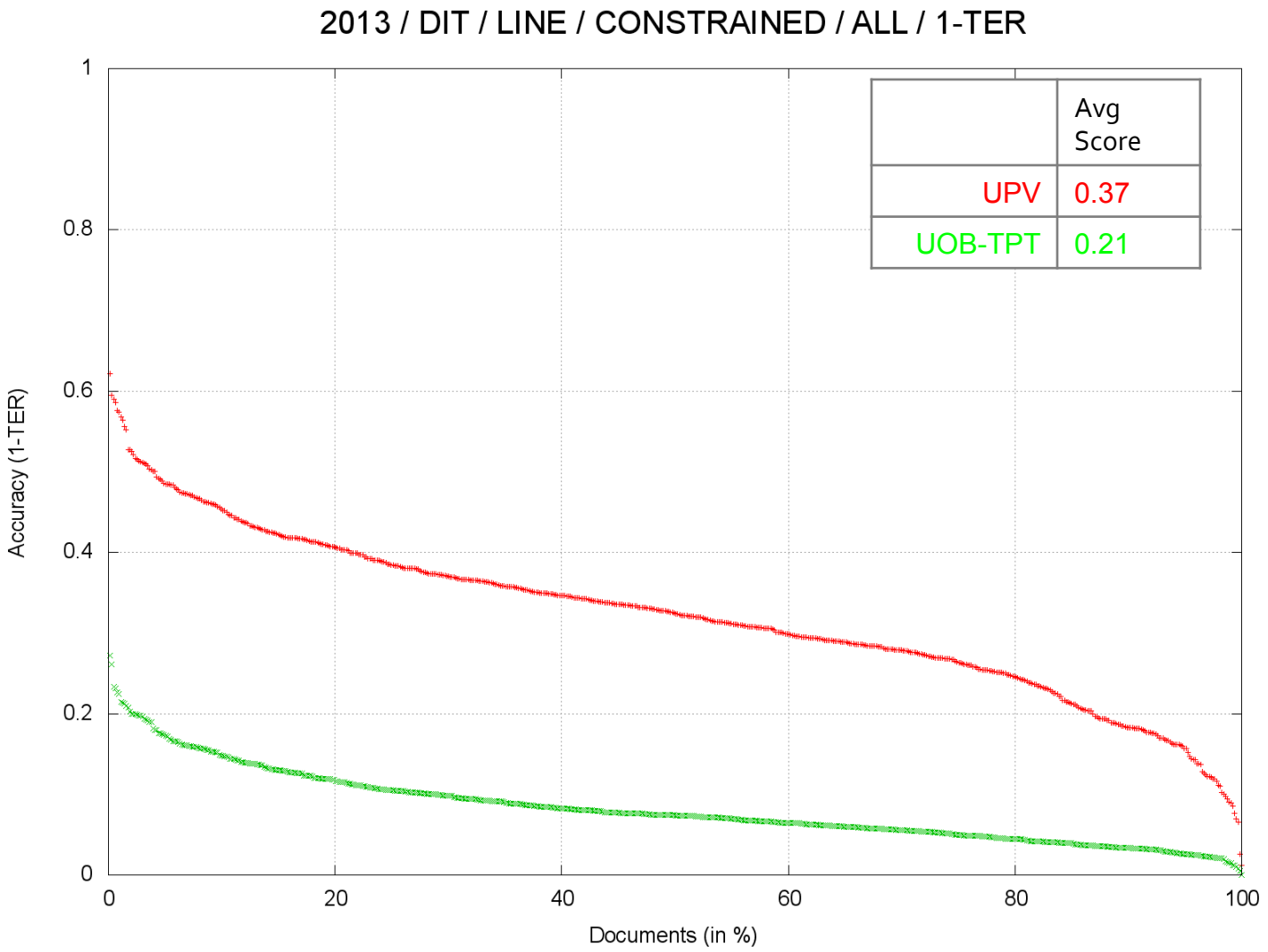
# Results: DIT Constrained Training

## Document Image Translation (OCR+MT)



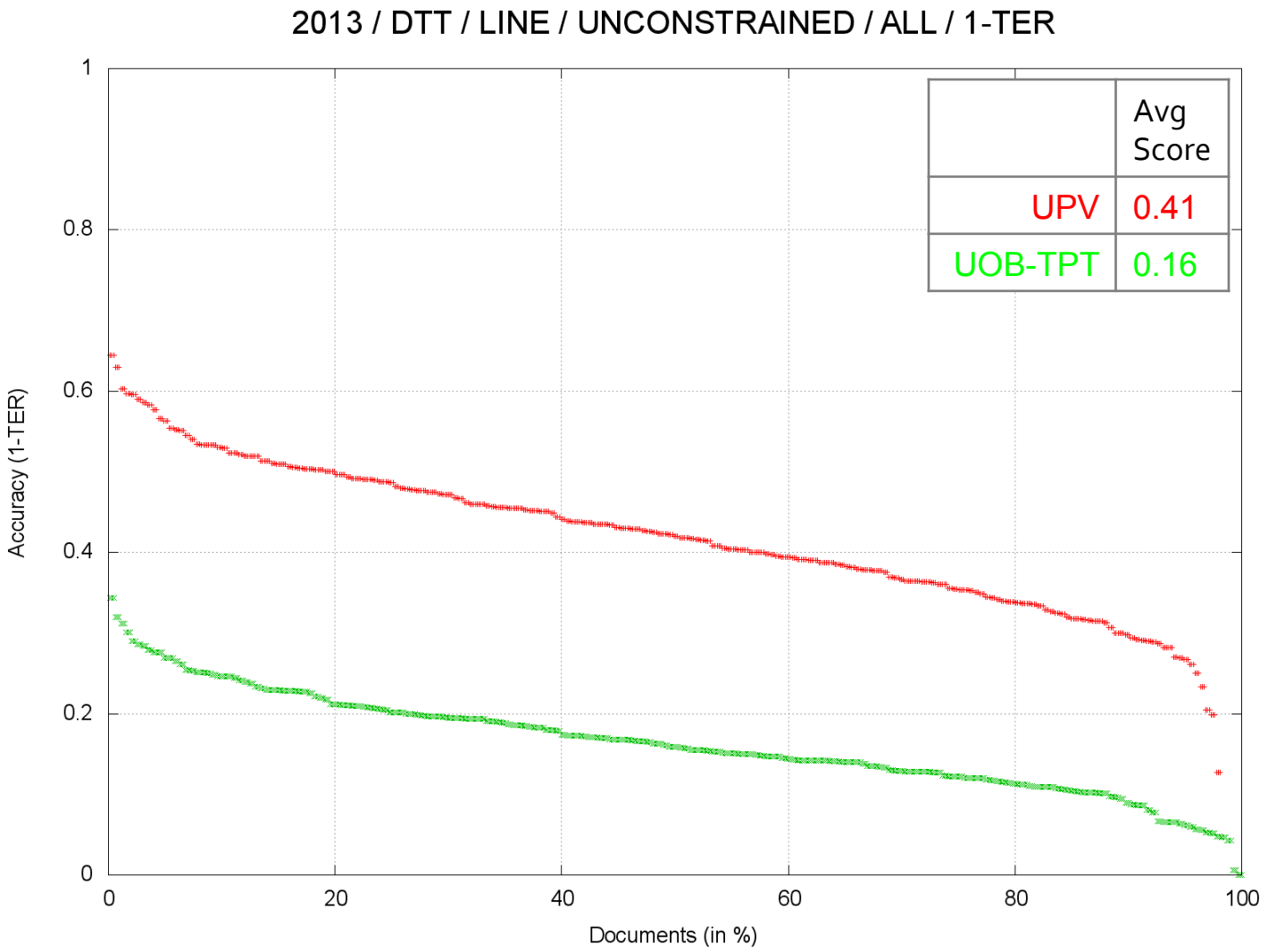
# Results: DTT Constrained Training

## Document Text Translation (MT)



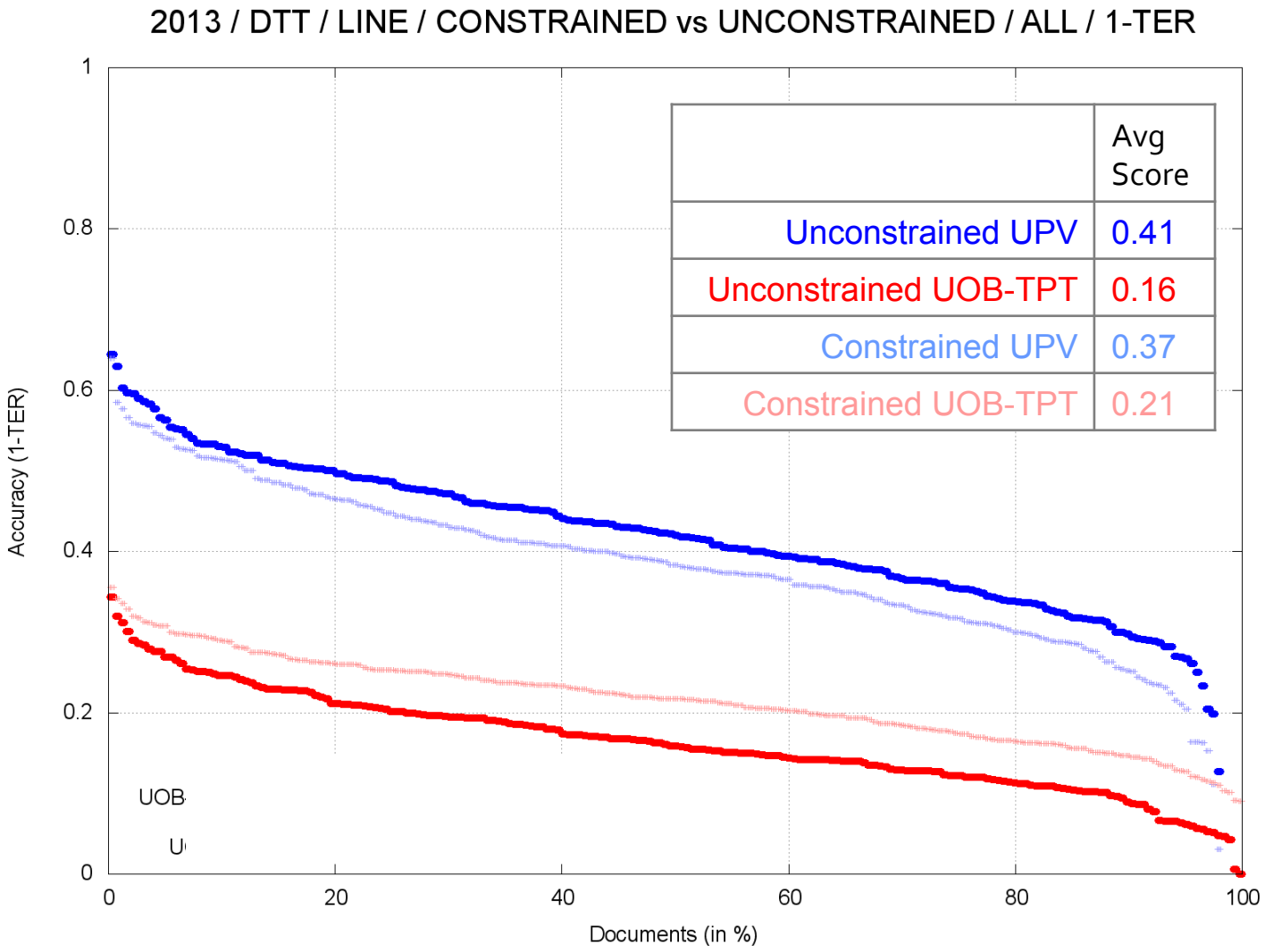
# Results: DTT Unconstrained Training

## Document Text Translation (MT)

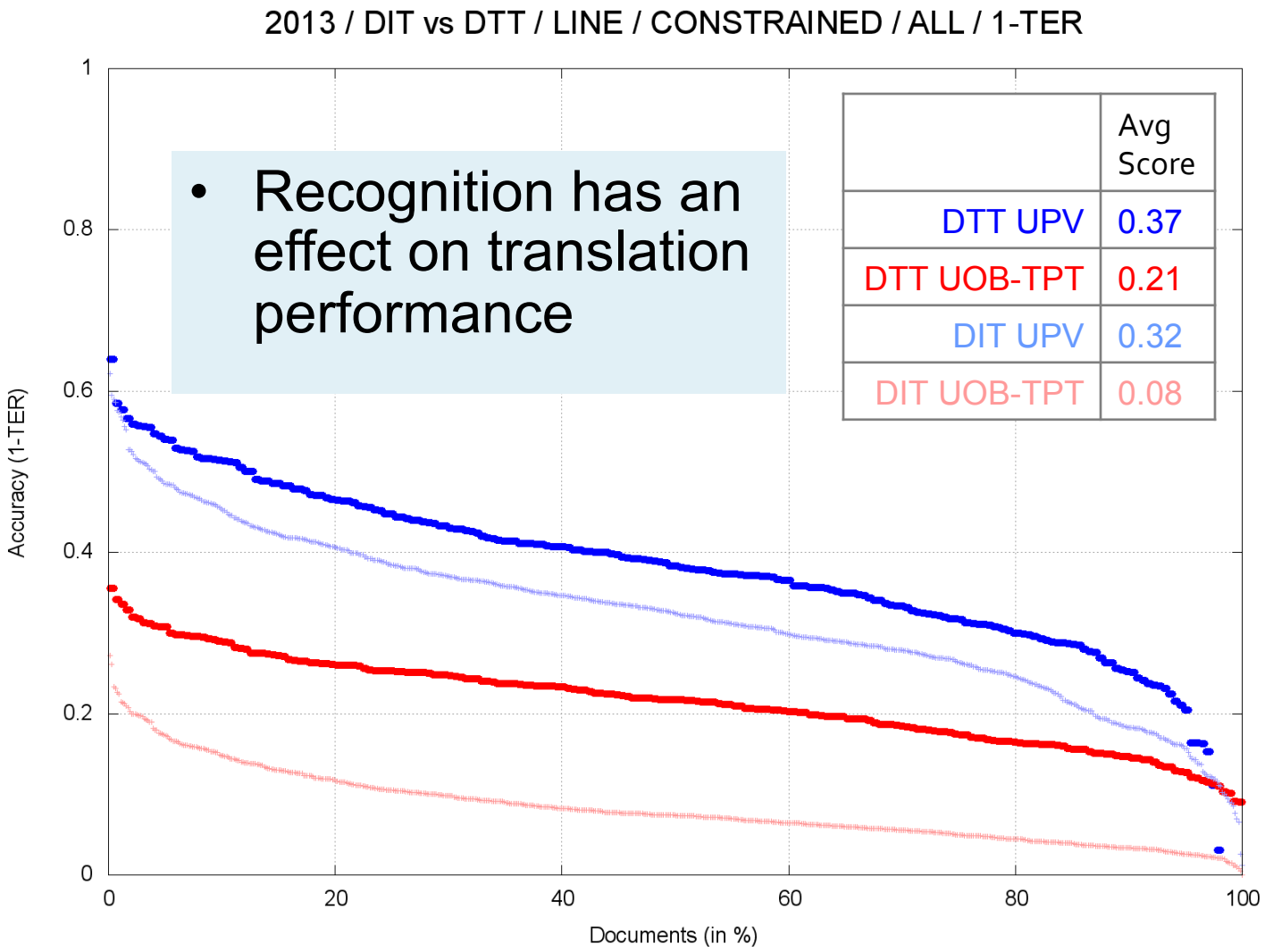


# Results: DTT Constrained vs Unconstrained

## Document Text Translation (MT)



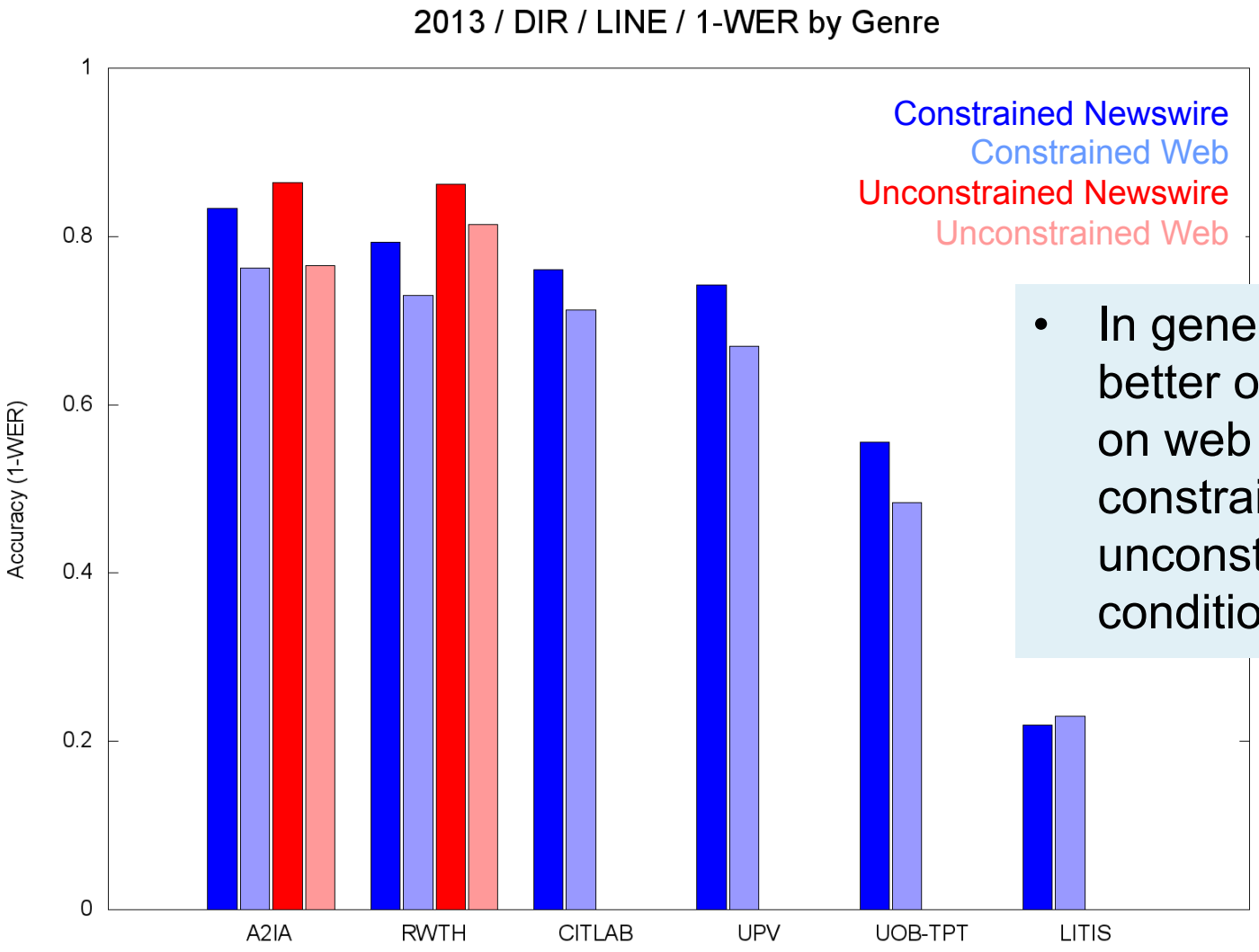
# Results: DIT vs DTT Constrained Training




- Recognition has an effect on translation performance

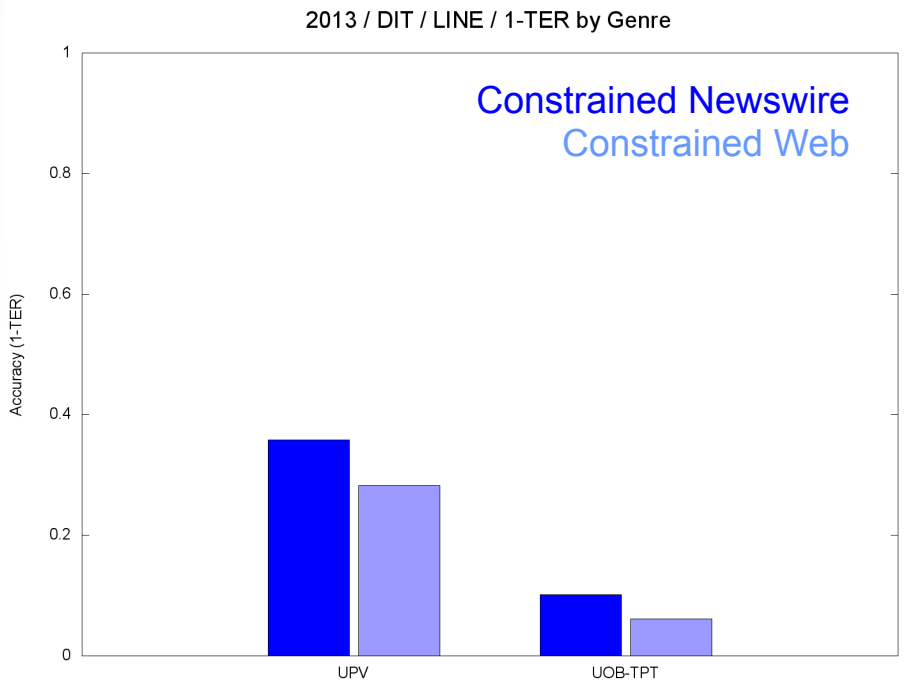
| | Avg Score |
|---|---|
| DTT UPV | 0.37 |
| DTT UOB-TPT | 0.21 |
| DIT UPV | 0.32 |
| DIT UOB-TPT | 0.08 |

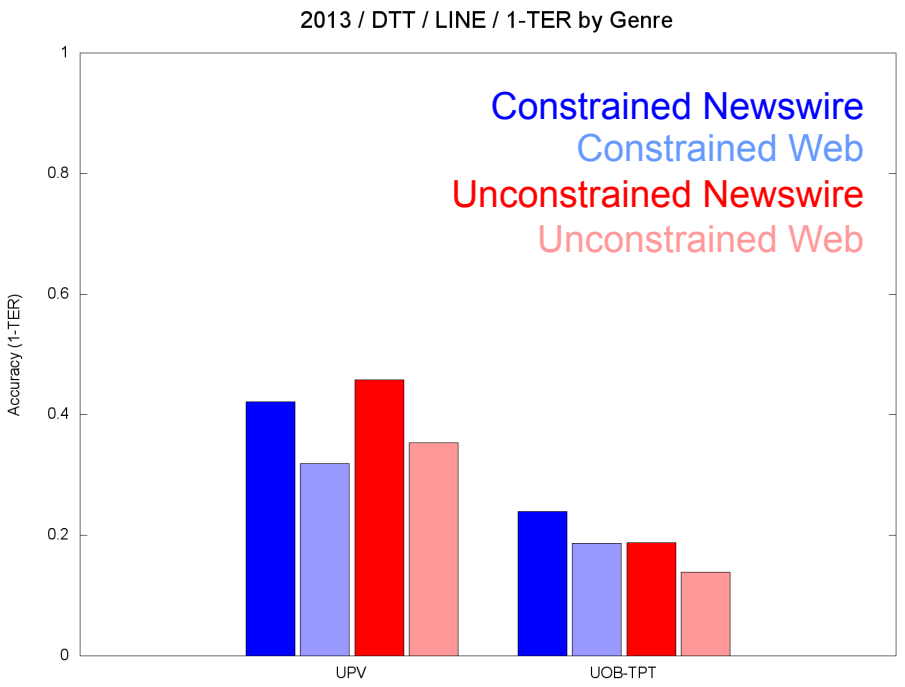
# Results: DIR Newswire vs Web




- In general systems did better on newswire than on web data for both constrained and unconstrained training condition

# Results: Translation Tasks Newswire vs Web



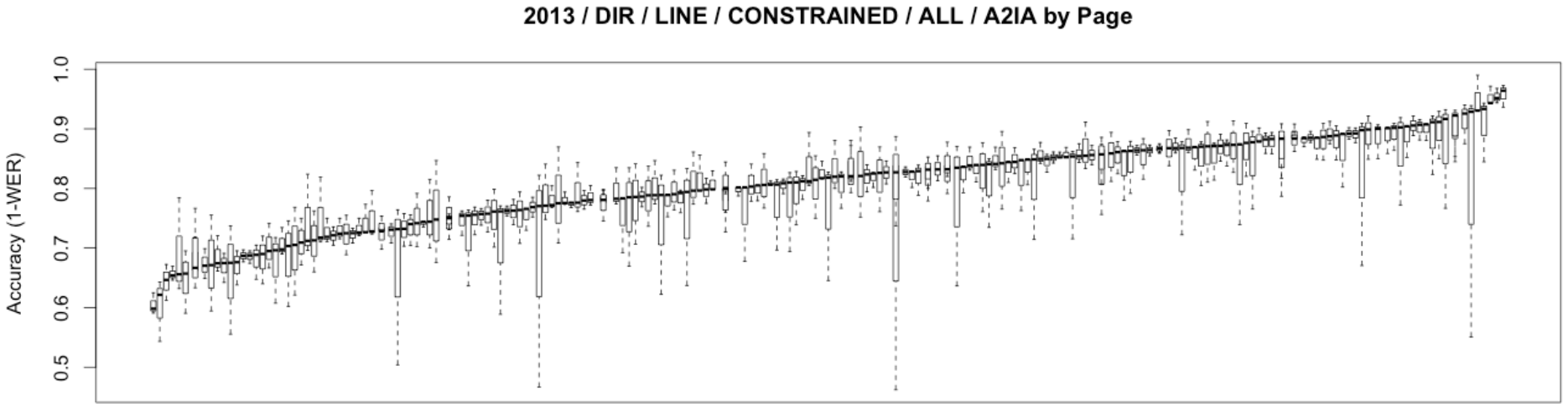



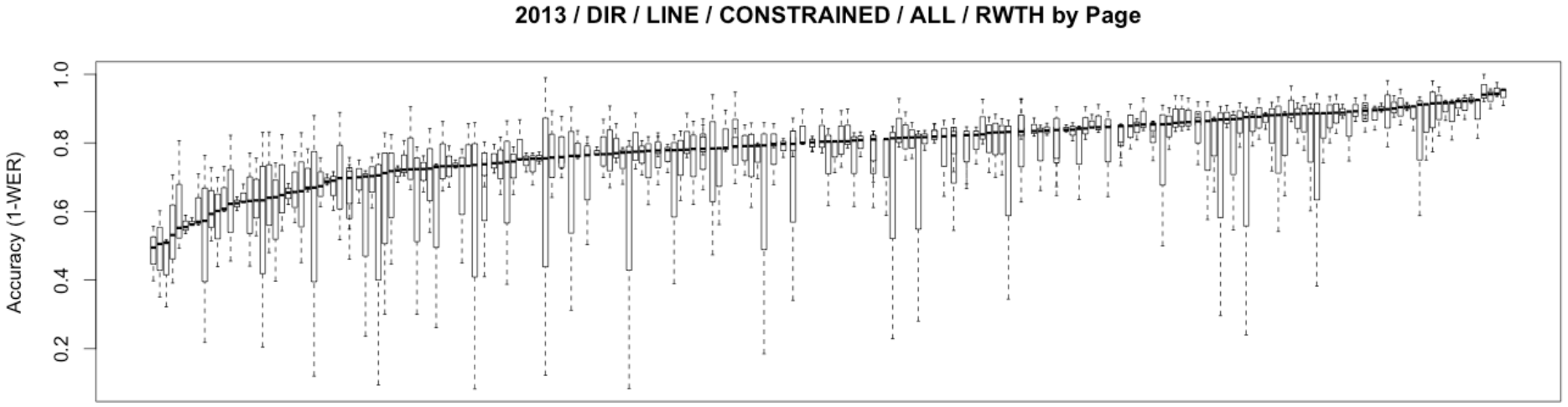
# Results: Translation Tasks by Metric

| Task | Training Condition | Team ID | TER | BLEU | METEOR |
|---|---|---|---|---|---|
| DIT | Constrained | UPV | 1 | 1 | 1 |
| | | UOB-TPT | 2 | 2 | 2 |
| DTT | Constrained | UPV | 1 | 1 | 1 |
| | | UOB-TPT | 2 | 2 | 2 |
| | Unconstrained | UPV | 1 | 1 | 1 |
| | | UOB-TPT | 2 | 2 | 2 |

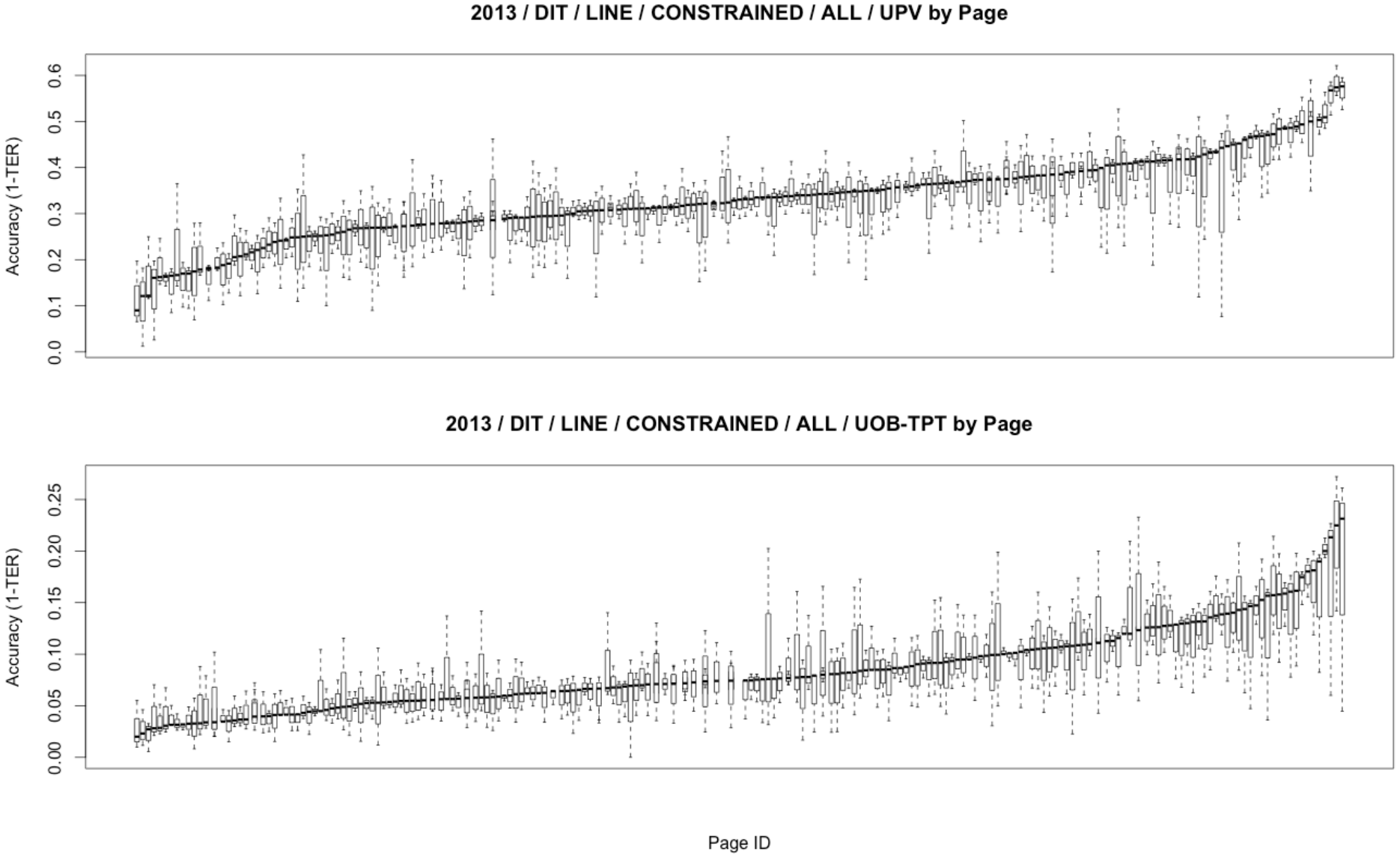
- Different metrics give the same ranking for these two systems

# Results: Scribe Effect DIR Constrained

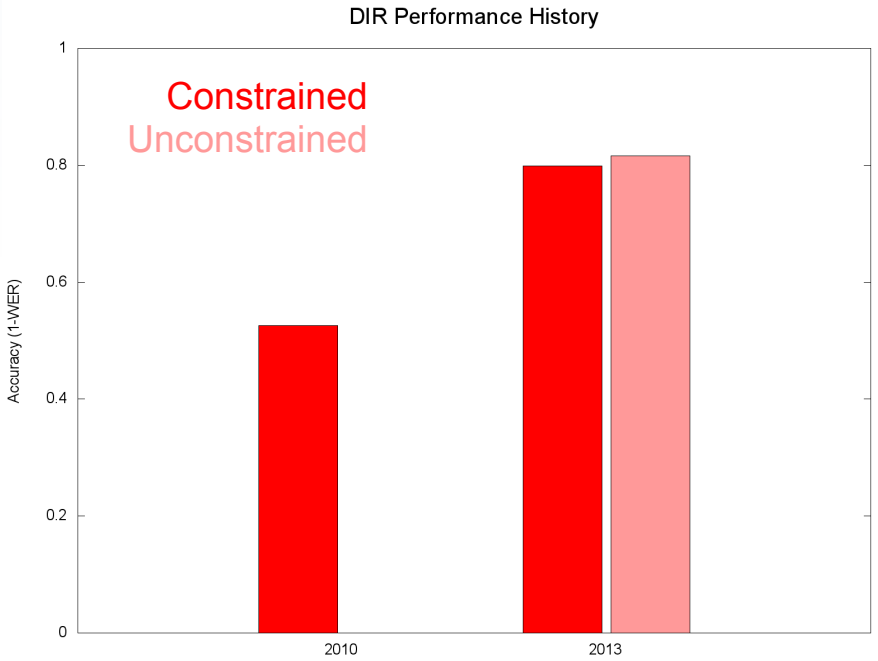


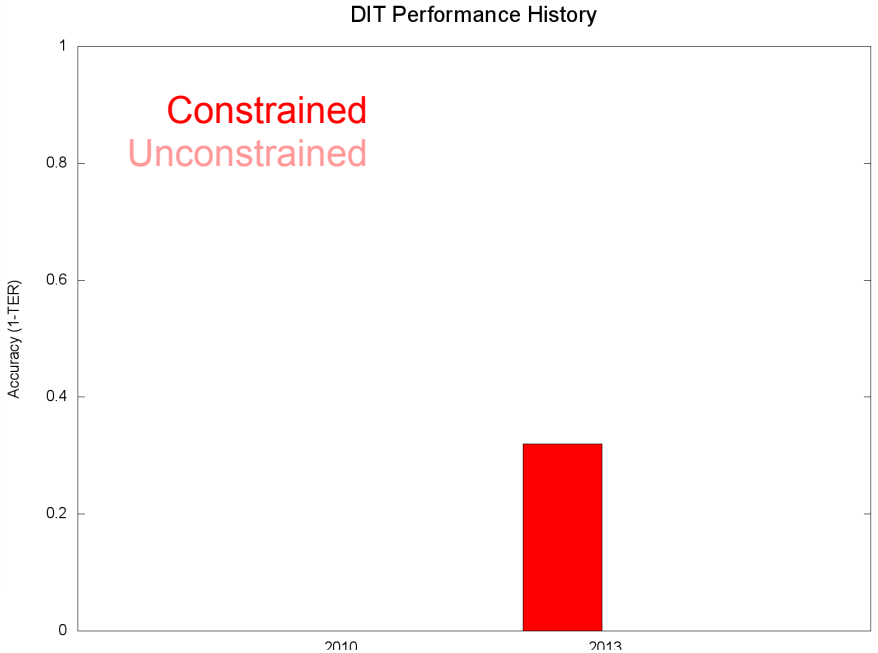



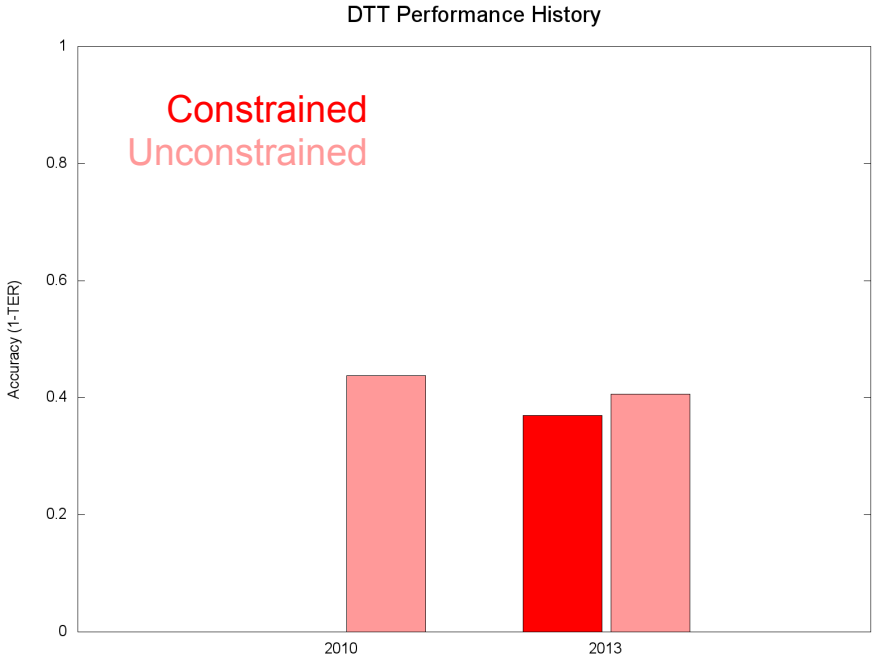
# Results: Scribe Effect DIT Constrained



# Results: 2013 vs 2010



# Summary

- Web data is harder than newswire data for recognition and translation tasks
  - Suggest data domain is an important factor in performance
- Scribe has an effect on recognition performance
- General improvement seen in recognition performance 2013 compared to 2010
  - Consider using a progress test set to keep scribe and document effect constant from year to year
  - Different participants in different year making progress tracking difficult
- Low participation and even lower for translation tasks
  - Is it because not much interest?
  - Should we move to other tasks?

# Thank You!

- DARPA
- The evaluation participants
- My workshop co-organizers
  - Volker Maergner
  - Haikal El Abed
  - Apostolos Antonacopoulos

# Invitation to 11th IAPR International Workshop on Document Analysis Systems

- Purpose:
  - Assemble industry, academic and government researchers to discuss many aspects of document analysis systems
- When: April 7th-10th, 2014
- Where: Tours – Loire Valley, France
- Important dates:
  - Regular paper submission: Sept 30, 2013
  - Short paper submission: Dec 20, 2013
  - Early Registration: Feb 9, 2014
- More information:
  - http://das2014.sciencesconf.org