

# **MET- 2013**

**Komal Agarwal**

# Introduction

## Morpheme Extraction Task

- Task held second time after introducing in FIRE-2012
- The task was closely modeled based on the Morphochallenge 2010 conducted by Department of Computer Science, Aalto University, Finland.  
[<http://research.ics.aalto.fi/events/morphochallenge2010/>]
- Task offered in five languages - Bengali, Gujarati, Hindi, Marathi and Odia
- Linguistic evaluation introduced this year for Tamil and Bengali

# Objective

- Objective is to encourage development of systems which discover morphemes in Indian languages
- Many Indian Languages are morphologically rich
- Morphological Analysis hence important for IR experiments in Indian Languages

# List of Participants

<b>S. No</b>	<b>Institute</b>	<b>Participant Names</b>	<b>Language(s) supported</b>
1	ISM-Dhanbad	Amit Jain, Nitish Gupta, Sukomal Pal	language independent
2	AUKBC, Chennai	Sobha Lalitha Devi, Marimuthu K, Vijay Sundar Ram R, Bakiyavathi T, Amudha K	Tamil
3	MANIT, Bhopal	Anubha Jain, Sujoy Das	Hindi

# The Task

The participating systems were asked to submit their Morpheme extraction systems.

System should be such that:

Input - large lexicon (already provided as test data to the participants)

Output - bicolon file containing tab separated word <tab> morpheme  
<space> affix grammar

# Evaluation methodology

## Runs-Information for IR Evaluation:

- Terrier-3.5
- Corpora, Queries, Qrels: Adhoc FIRE 2011
- Ranking model: In\_expC2
- Stopwords: FIRE data
- TrecQuery tags: TITLE,DESC (T,D)

## Evaluation:

Trec-Eval-9.0 (Metric used: MAP)

# Evaluation methodology

## Linguistic Evaluation methodology:

- For Tamil and Bengali languages
- Sampled pairs of words from the proposed morpheme analyses of the systems were compared against sampled pairs of words from the gold standard data
- If sampled from gold standard data it is called Recall
- When sampled from proposed analysis, it is called Precision
- F-measure is computed as final score
- Code from Morphochallenge 2010 has been used

<http://research.ics.aalto.fi/events/morphochallenge2010/>

# MET Results

ISM:

Language	Baseline MAP	MAP obtained	% improvement
Bengali	0.2740	0.3158	<b>15.25%</b>
Hindi	0.2821	0.2793	-0.99%
Gujarati	0.2677	0.2824	5.49%
Marathi	0.2320	0.2797	<b>20.56%</b>
Odia	0.1537	0.1583	2.99%

# MET Results

MANIT:

Language	Baseline MAP	MAP obtained	% improvement
Hindi	0.2821	0.2917	3.40%

# Results analysis

Comparison with 2012 scores:

Language	2012's best MAP	compare	2013's best MAP
Bengali	0.3307	>	0.3158
Hindi	0.2963	>	0.2917
Gujarati	0.2824	=	0.2824
Marathi	0.2797	=	0.2797
Odia	0.1537	<	0.1583

# MET Results

Linguistic Evaluation:

System	Language	Precision	Recall	F-measure
AUKBC	Tamil	<b>84.29%</b>	<b>88.15%</b>	<b>86.17%</b>
ISM	Tamil	<b>80.22%</b>	18.86%	30.54%
ISM	Bengali	<b>60.64%</b>	32.15 %	42.02%

**Numbers in bold show significant accuracy rate**

# Conclusion

- 
- Encouraged participants to experiment and improve their systems and obtain better scores.
- Most of the systems show significant improvement over baseline scores
- Good gold standard data are required in all languages for better linguistic evaluation.

# Acknowledgements

- Prof. Prasenjit Majumder (DA-IICT)
- Rashmi Sankepally and Parth Mehta
- Prof. Mandar Mitra (ISI, Kolkata)
- All the organizers of FIRE-2012
- All the esteem research personnel who have contributed to the FIRE data
- All the participants of MET
- The developers of Terrier (Univ of Glasgow)

Thank you!