फूलदार वनस्पति, फूल

flower

সপুষ্পক , ফুলযুক্ত উদ্ভিদ

பூச்செடி

fiore

ફૂલછોડ , ફૂલદાર વનસ્પતિ

പൂച്ചെടി , പൂക്കളുള്ള ചെടി

ਫੁੱਲਦਾਰ ਬਨਸਪਤੀ , ਫੁੱਲ

花

फूलदार वनस्पति, फूल

flower

সপুষ্পক , ফুলযুক্ত উদ্ভিদ

பூச்செடி

fiore

ફૂલછોડ , ફૂલદાર વનસ્પતિ

പൂച്ചെടി , പൂക്കളുള്ള ചെടി

ਫੁੱਲਦਾਰ ਬਨਸਪਤੀ , ਫੁੱਲ

花

फूलदार वनस्पति, फूल

flower

সপুষ্পক , ফুলযুক্ত উদ্ভিদ

பூச்செடி

fiore

ફૂલછોડ , ફૂલદાર વનસ્પતિ

പൂച്ചെടി , പൂക്കളുള്ള ചെടി

ਫੁੱਲਦਾਰ ਬਨਸਪਤੀ , ਫੁੱਲ

花

फूलदार वनस्पति, फूल

flower

সপুষ্পক , ফুলযুক্ত উদ্ভিদ

பூச்செடி

fiore

ફૂલછોડ , ફૂલદાર વનસ્પતિ

പൂച്ചെടി , പൂക്കളുള്ള ചെടി

ਫੁੱਲਦਾਰ ਬਨਸਪਤੀ , ਫੁੱਲ

花

फूलदार वनस्पति, फूल

flower

সপুষ্পক , ফুলযুক্ত উদ্ভিদ

பூச்செடி

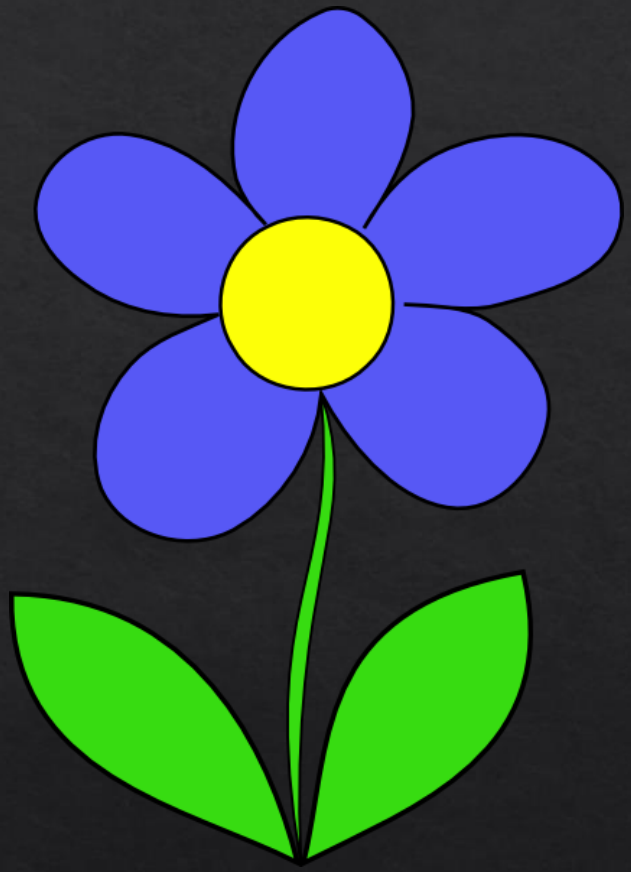
fiore

ફૂલછોડ , ફૂલદાર વનસ્પતિ

പൂച്ചെടി , പൂക്കളുള്ള ചെടി

ਫੁੱਲਦਾਰ ਬਨਸਪਤੀ , ਫੁੱਲ

花



A picture is worth a thousand words: Using OpenClipArt library to enrich IndoWordNet

Diptesh Kanojia, Shehzaad Dhuliawala, and Pushpak Bhattacharyya

Center for Indian Language Technology,
IIT Bombay

8th Global Wordnet Conference, Bucharest, 2016.

Roadmap

- Introduction
- Motivation
- Previous Work
- Architecture
- Retrieval & Scoring
- Results
- Manual Evaluation & Analysis
- Summary

Introduction

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.
- Many possible Sources for such a resource were considered including crawling data from Google based on Synset word query.
- Such crawls may have gotten us images from various sources, which may / may not have been under copyright.
- We chose OpenClipArt Library (OCAL) for the initial phase of the project to avoid any such complications.
- We managed to retrieve, rank and manually annotate a number of images for our purpose.
  - Analysis of the erroneous retrieval or linkages was done.

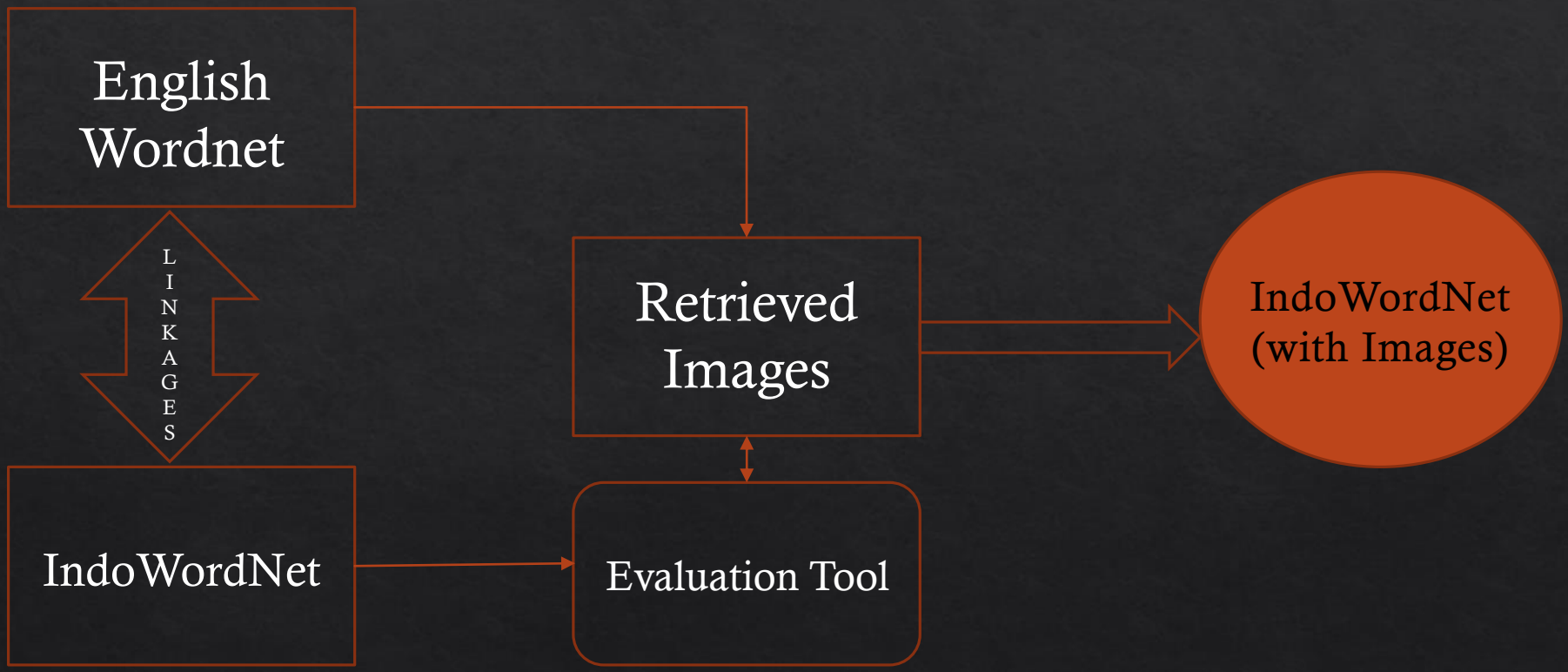
Motivation

- "Digital India" Initiative by Hon'ble Prime Minister Shri Narendra Modi.
- Use of WordNets for educations / in the education sector.
  - WordNet should be introduced in schools, thus, helping students and language enthusiasts alike.
- The English – Hindi pivot of IndoWordnet allows us to search for (in English) and tag (for all the Indian Languages)
- It is widely believed children grasp / learn better if visual aids are used. It is one of the three basic teaching and learning styles.
- Languages – so many languages in India!
- Language Learner problems:
  - Finding the same concept in other language: Concept / Synset mapping in Wordnets.
  - Script readability of a user.

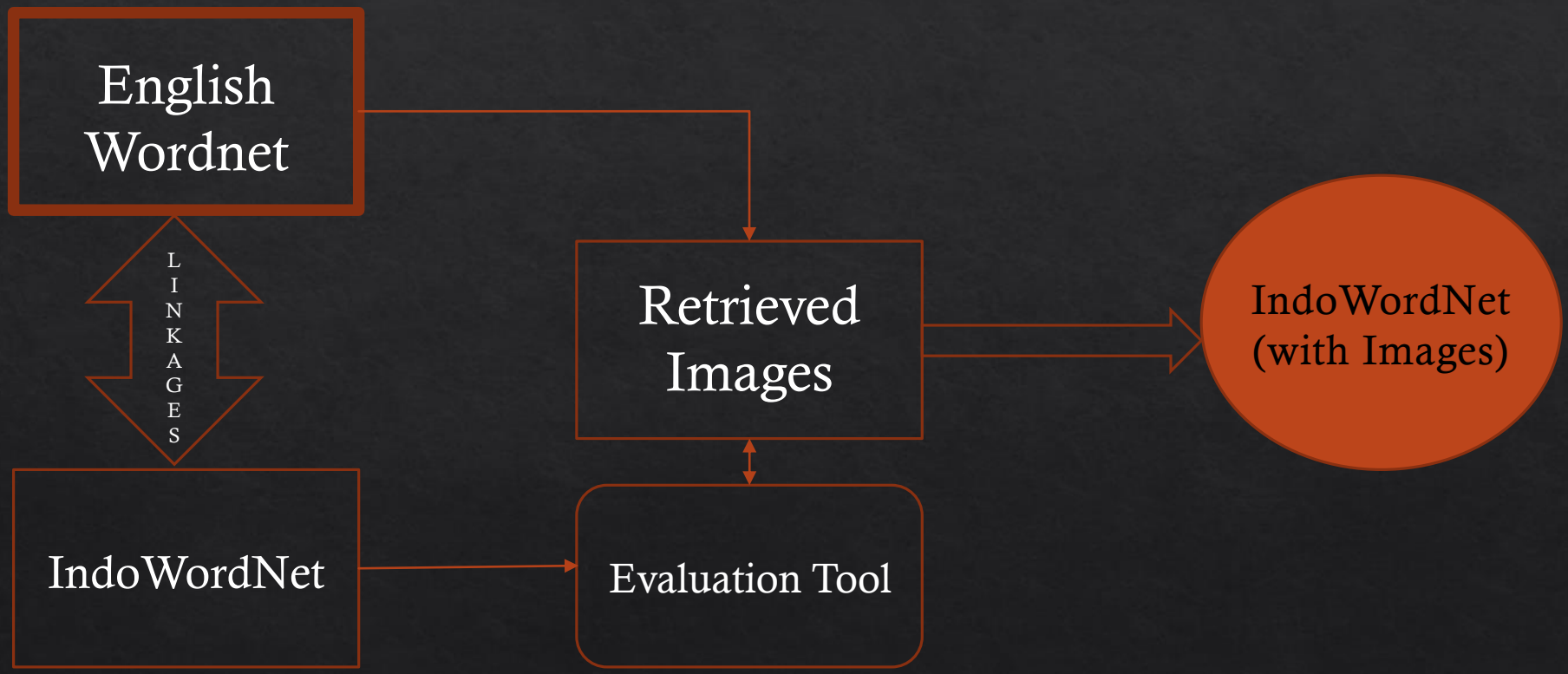
Related Work

- Bond et al. (2009) used OCAL to enhance the Japanese WordNet.
  - They were able to mine 874 links for 541 synsets.
- Imagenet (Deng et al., 2009) is a similar project for Princeton WordNet which provides images/URLs for a concept. It contains 21,841 synsets indexed with 14,197,122 images.
  - Imagenet uses computer vision to recognize the objects in an image, and provides tags / bag of words to the images.
- We use a rather simple alternative, and use OCAL to retrieve images and rank them.

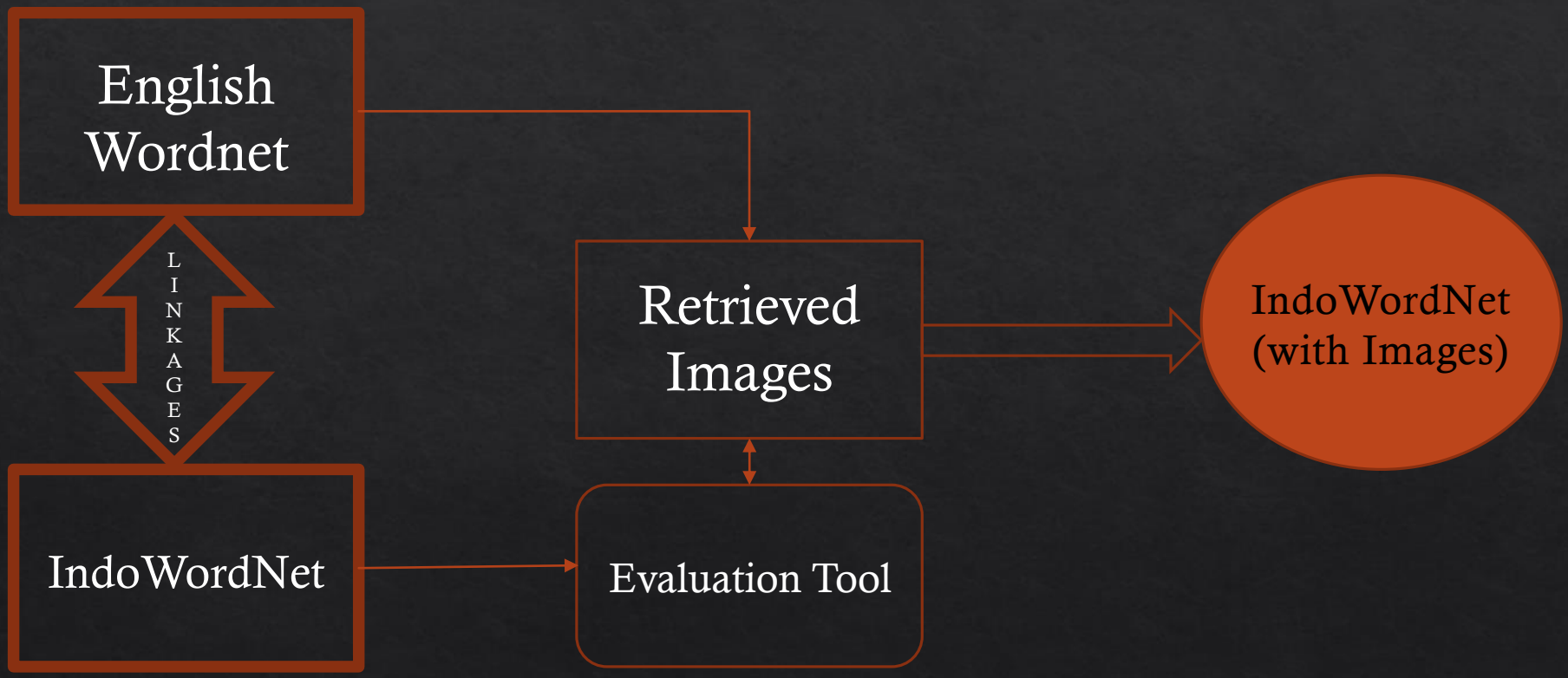
Architecture



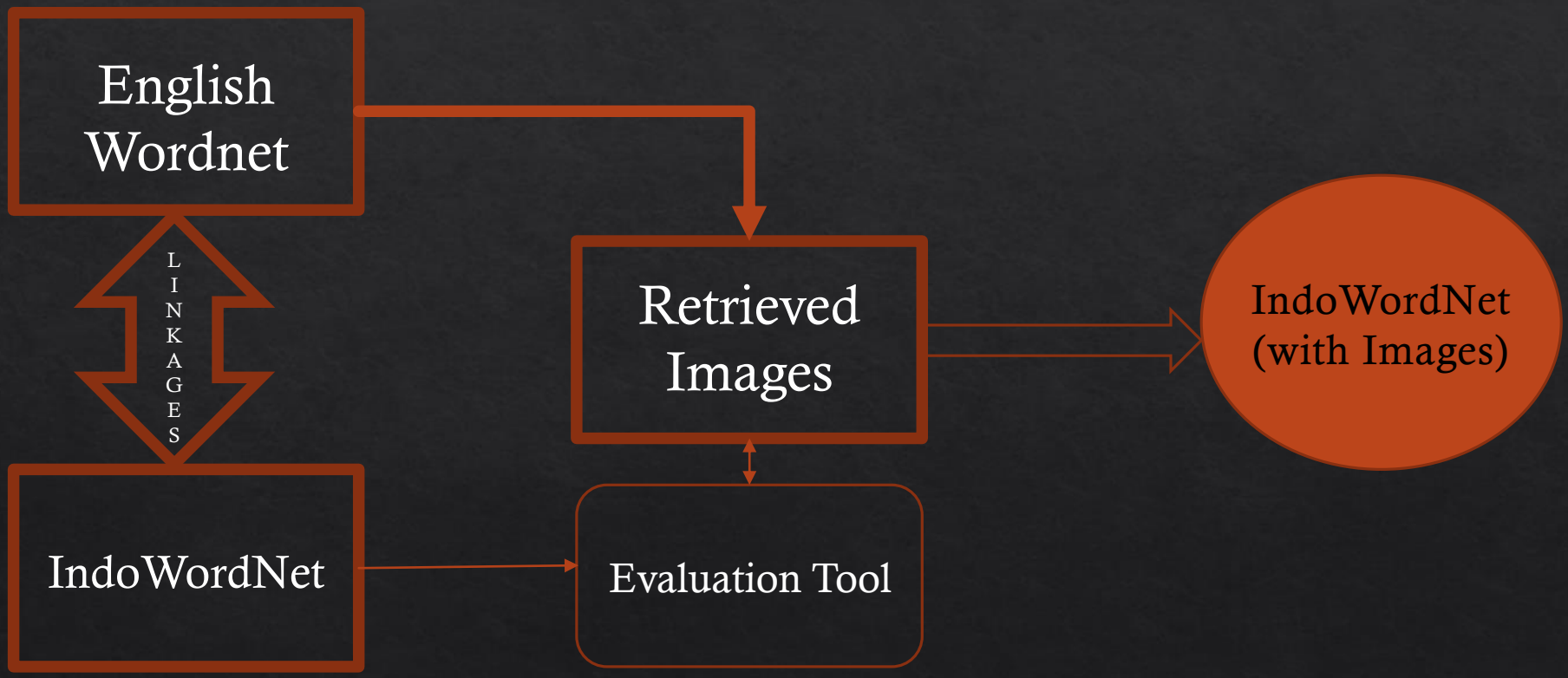
Architecture



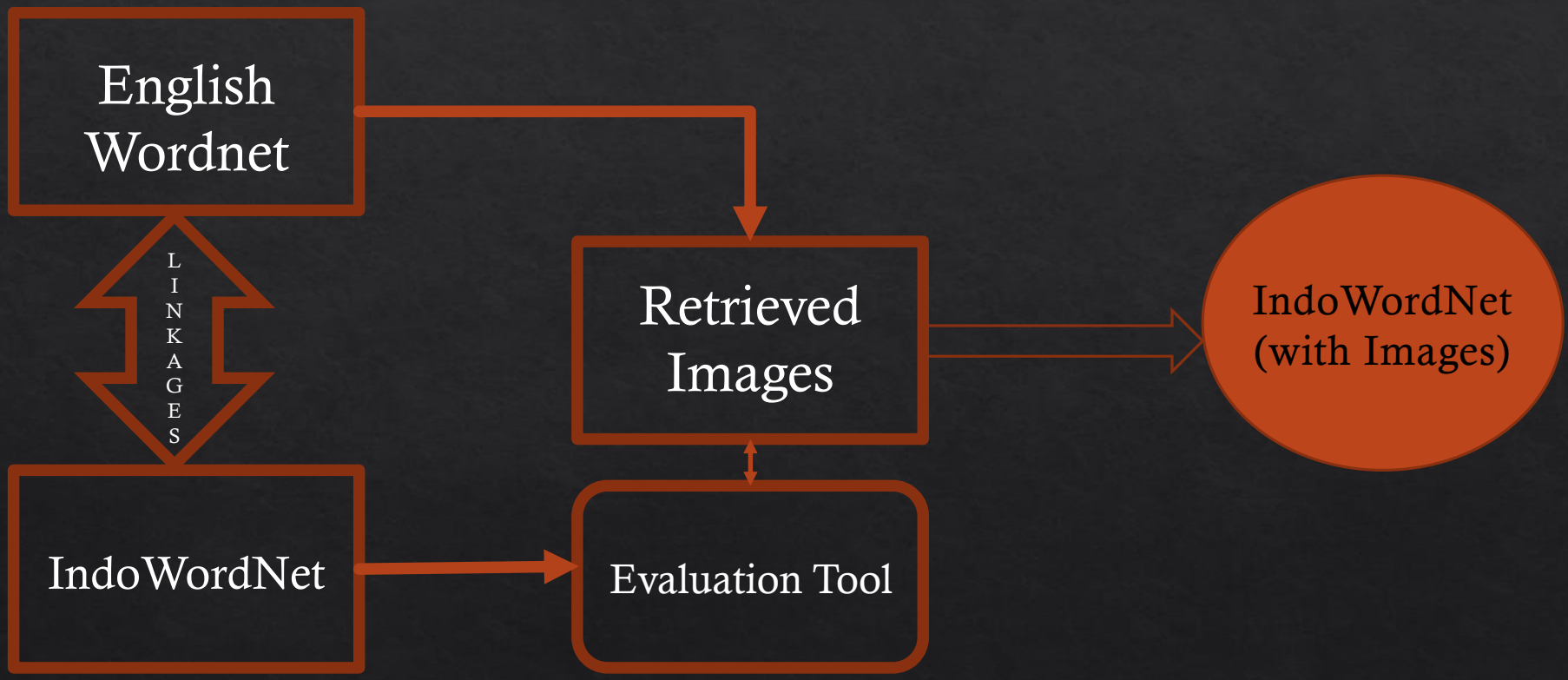
Architecture



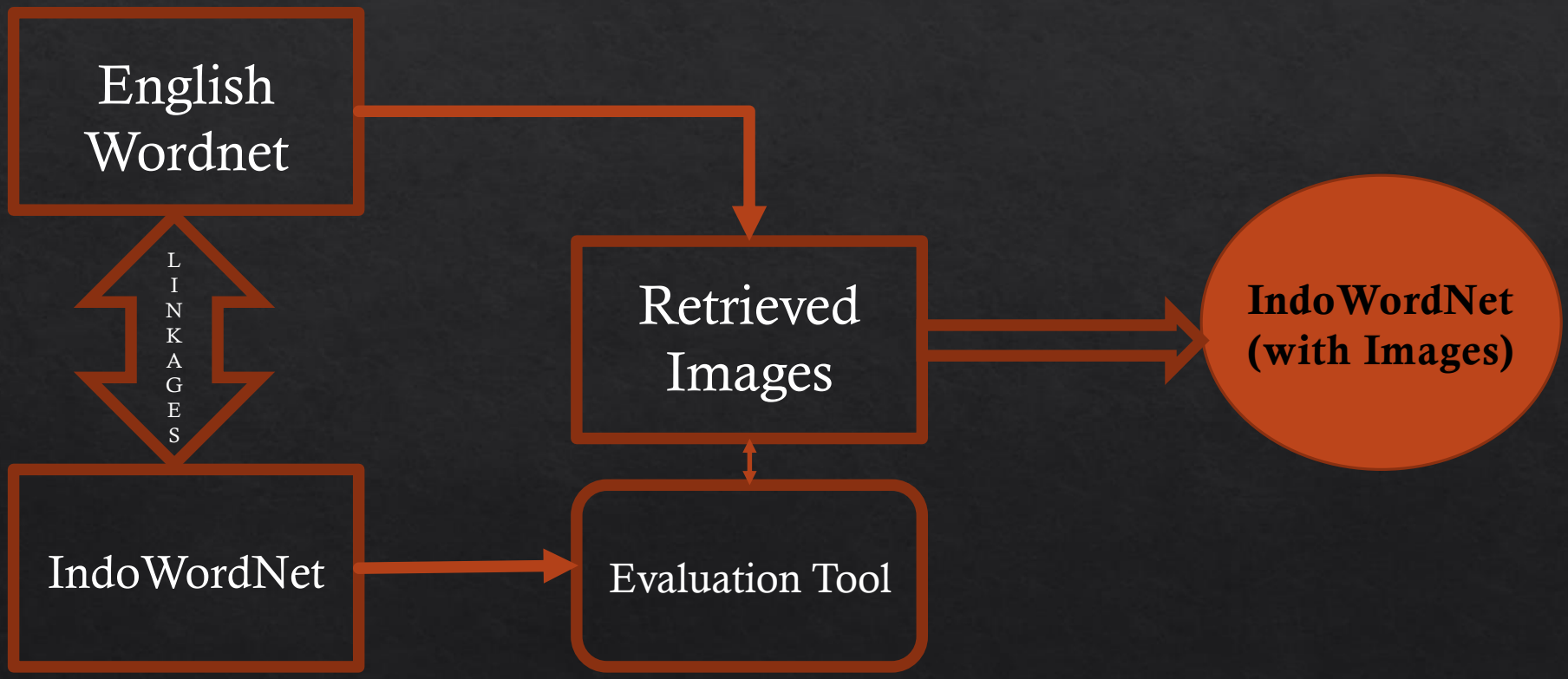
Architecture



Architecture



Architecture



Retrieval & Scoring

- For each synset, the head word was used as the query of the OCAL API.
- The resultant images' meta-tags were compared to the words in the definition and synset list.
  - A greater overlap provided the result a greater score while greater non-overlapping terms bore a negative cost.
- The results were ranked and the **top three** images were chosen.

Evaluation

- A PHP-MySQL based evaluation tool was created for an evaluator to manually evaluate the images.
  - Each image was evaluated on the basis of their Indian language synsets.

Evaluation


NONE
OF
THESE
wild_boar, boar, Sus_scrofa
जंगली सुअर,जंगली शूकर,जंगली सूअर,जंगली सूकर,वाराह,वराह,बराह,बाराह,दीर्घरद,वज्ररद,जंगली रेवट,वज्रदंत,वज्रदन्त
जंगलों में पाया जाने वाला सुअर जो अपने नुकीले दाँतों के लिए जाना जाता है
Submit and proceed
I can't decipher!

Evaluation

- A PHP based evaluation tool was created for an evaluator to manually evaluate the images.
  - Each image was evaluated on the basis of their Indian language synsets.
  - Evaluation done for one language can be used for all – Hindi is the Pivot!

Result

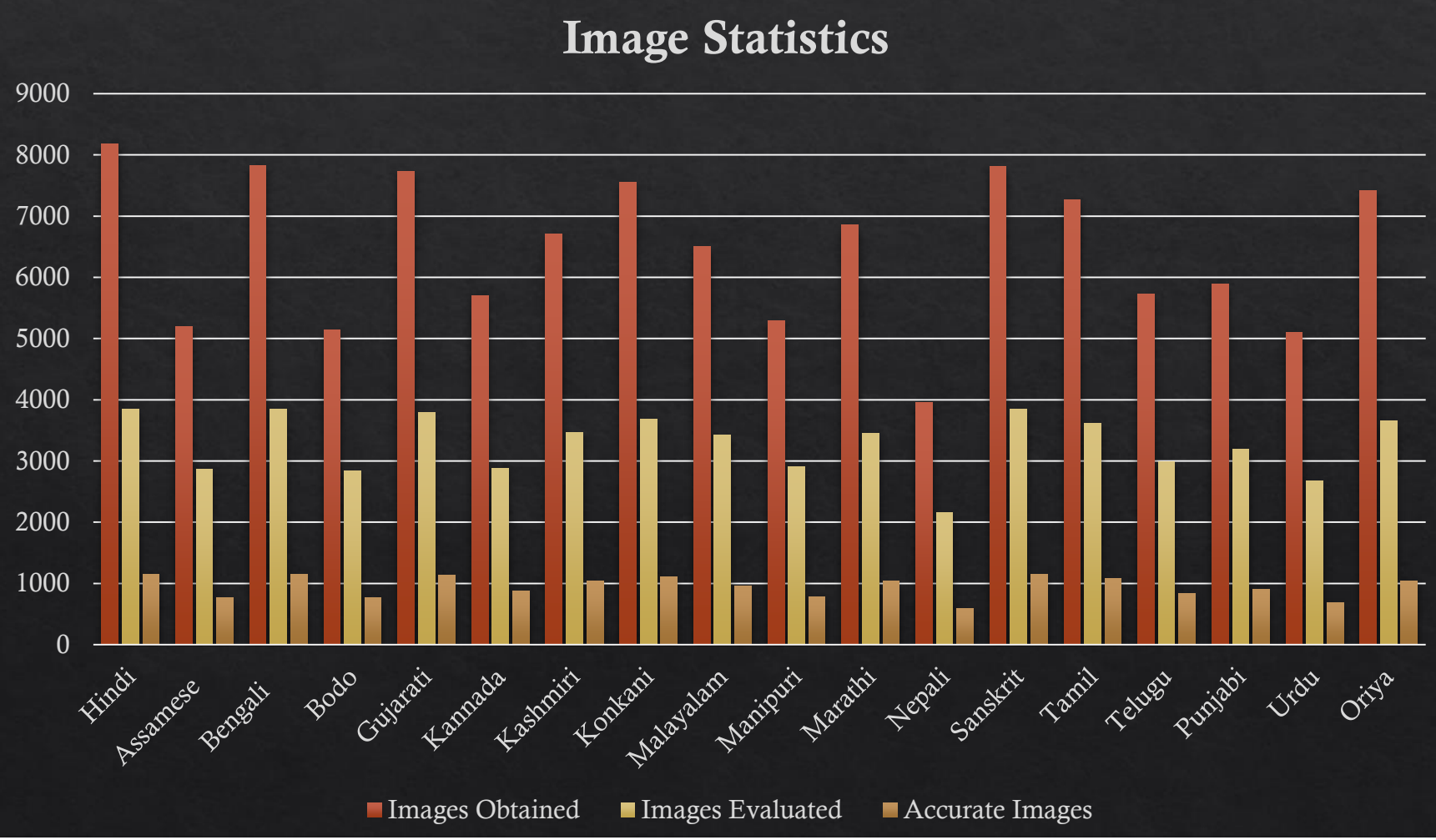
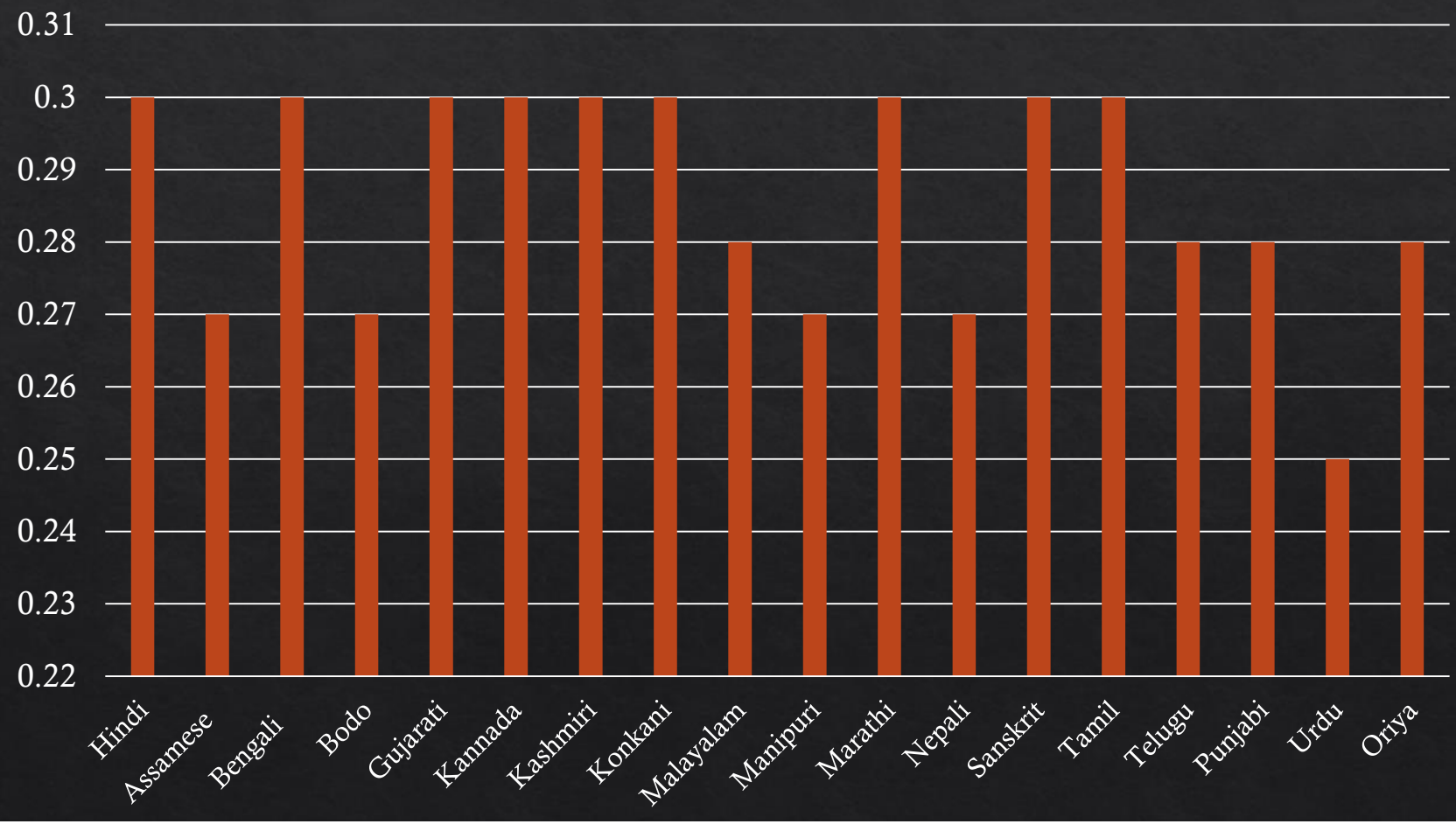


Image Statistics
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Hindi
Assamese
Bengali
Bodo
Gujarati
Kannada
Kashmiri
Konkani
Malayalam
Manipuri
Marathi
Nepali
Sanskrit
Tamil
Telugu
Punjabi
Urdu
Oriya
Images Obtained
Images Evaluated
Accurate Images

Result


Precision Statistics
0.31
0.3
0.29
0.28
0.27
0.26
0.25
0.24
0.23
0.22
Hindi
Assamese
Bengali
Bodo
Gujarati
Kannada
Kashmiri
Konkani
Malayalam
Manipuri
Marathi
Nepali
Sanskrit
Tamil
Telugu
Punjabi
Urdu
Oriya

Qualitative Analysis

- **No images found (3,390 / 11,573)**

Qualitative Analysis

- No images found (3,390 / 11,573)
  - Abstract Nouns

Qualitative Analysis

- **No images found (3,390 / 11,573)**
  - **Abstract Nouns**
    - "गुलछर्रा" (*"gUlchharra"*) - which translates to "profligacy, extravagance"

Qualitative Analysis

- **No images found (3,390 / 11,573)**
  - **Abstract Nouns – too vauge for an image to do justive to a concept.**
    - "गुलछर्रा" (*"gUlchharra"*) - which translates to "profligacy, extravagance"

Qualitative Analysis

- Images Found (8,183 / 11,573)

Qualitative Analysis

- Images Found (8,183 / 11,573)
  - **Common Nouns**

Qualitative Analysis

- Images Found (8,183 / 11,573)
  - **Common Nouns**
    - "मोमबत्ती" (“momBatti”) - meaning “candle”

Qualitative Analysis

- Images Found (8,183 / 11,573)
  - **Common Nouns**
    - "मोमबत्ती" (“momBatti”) - meaning “candle”

Qualitative Analysis

- Images Found (8,183 / 11,573)
  - **Common Nouns**
    - "मोमबत्ती" (“momBatti”) - meaning “candle”
    - "मस्जिद" (*“maszid”*) - meaning “mosque”

Qualitative Analysis

- Images Found (8,183 / 11,573)
  - **Common Nouns**
    - "मोमबत्ती" (“momBatti”) - meaning “candle”
    - "मस्जिद" (*“maszid”*) - meaning “mosque”

Summary

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.

Summary

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.
- IndoWordNet has 11,573 direct linkages, and 8184 hypernymy linkages. We use only directly linked Noun concepts to mine OCAL.

Summary

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.
- IndoWordNet has 11,573 direct linkages, and 8184 hypernymy linkages. We use only directly linked Noun concepts to mine OCAL.
- We used OpenClipArt Library (OCAL) API to retrieve images for a head word from a synset, and ranked them based on a naïve algorithm.

Summary

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.
- IndoWordNet has 11,573 direct linkages, and 8184 hypernymy linkages. We use only directly linked Noun concepts to mine OCAL.
- We used OpenClipArt Library (OCAL) API to retrieve images for a head word from a synset, and ranked them based on a naïve algorithm.
- An evaluation tool was created, and used to evaluate more than three thousand synset images (a total of nine thousand images).

Summary

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.
- IndoWordNet has 11,573 direct linkages, and 8184 hypernymy linkages. We use only directly linked Noun concepts to mine OCAL.
- We used OpenClipArt Library (OCAL) API to retrieve images for a head word from a synset, and ranked them based on a naïve algorithm.
- An evaluation tool was created, and used to evaluate more than three thousand synset images (a total of nine thousand images).
- We record a maximum precision value of 0.3 for Hindi, and a minimum of 0.25 for Urdu.

Summary

- **Goal:** To enrich the semantic lexicon of various Indian Languages by mapping it with images.
- IndoWordNet has 11,573 direct linkages, and 8184 hypernymy linkages. We use only directly linked Noun concepts to mine OCAL.
- We used OpenClipArt Library (OCAL) API to retrieve images for a head word from a synset, and ranked them based on a naïve algorithm.
- An evaluation tool was created, and used to evaluate more than three thousand synset images (a total of nine thousand images).
- We record a maximum precision value of 0.3 for Hindi, and a minimum of 0.25 for Urdu.
- We continue to evaluate, and find better methods for retrieving images.

Acknowledgements

Acknowledgements

- Rajita Shukla

Acknowledgements

- Rajita Shukla
- Jaya Jha

Acknowledgements

- Rajita Shukla
- Jaya Jha
- Laxmi Kashyap

Acknowledgements

- Rajita Shukla
- Jaya Jha
- Laxmi Kashyap
- Meghna Singh

Acknowledgements

- Rajita Shukla
- Jaya Jha
- Laxmi Kashyap
- Meghna Singh
- Ankit

Acknowledgements

- Rajita Shukla
- Jaya Jha
- Laxmi Kashyap
- Meghna Singh
- Ankit
- Amisha

Acknowledgements

- Rajita Shukla
- Jaya Jha
- Laxmi Kashyap
- Meghna Singh
- Ankit
- Amisha
- Department of Electronics and Information Technology, Ministry of Communications and IT, Government of India.
- IRCC, IIT Bombay for the Travel Fund.

Thank you!

- Thank you for all your attention.
- Questions ?
- Suggestions ?