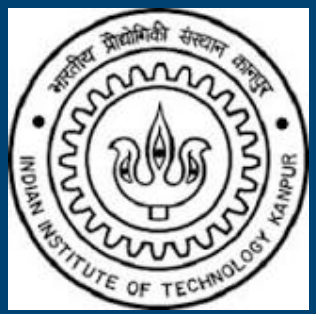
CS365 Project
LANGUAGE LEARNING FROM BENGALI COMMENTARY IN VIDEOS

Guide: Prof. Amitabha Mukherjee
Anusha Chowdhury



March 28, 2015

Problem Statement

The system takes a set of commentaries on videos as input for the purpose of **Bengali language learning**.
The main aim is **word learning** and **syntax learning**. Syntax basically refers to the positioning of subject, object and verb.
The input videos have agents performing some actions on coloured objects and there is a target.

Motivation

Main motivation was the urge to learn Bengali language from scratch just as a child does.

Figure : A child learns language by **semantic mapping**: how linguistic elements relate to visible situations

DYNAMIC NLP: Preliminaries

- Process of learning involves continuous expansion of the already acquired language model.
- There is continuous learning from new sentences
- All evaluations of sentence semantics is partial
- Associations between word or phrase and meaning is dynamic which may broaden or shrink.

Our Approach

- Collect the input data (Bengali commentary)
- Identify the **Family-lect** and the **Multi-lect** from the corpus
- Perform **contrastive association** based on probability calculations
- Words like 'and' are ignored to shorten down the corpus
- Patterns are learnt by applying **ADIOS** algorithm on the family lect after pruning the corpus
- Learning the sentence syntax from verb phrases
- **Morphosyntactic discovery**: Text-based morphological similarity analysis gives rise to clusters.

The video

A glimpse of the video, each having 16 frames:

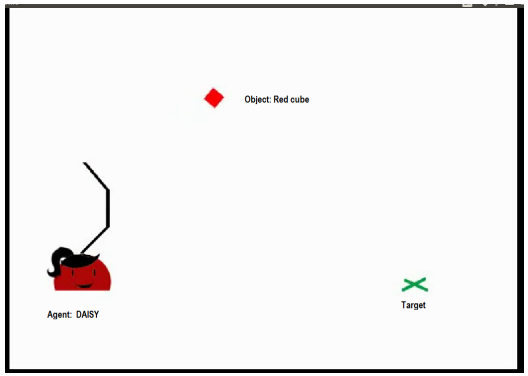


Figure : Daisy is throwing a Red Cube

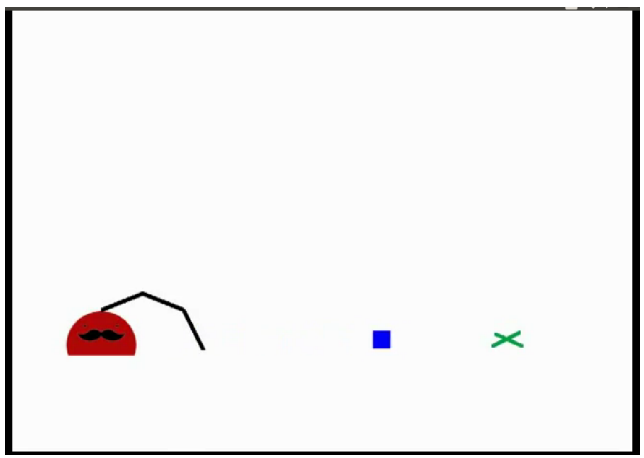


Figure : Dome is rolling a blue cube

Pre-linguistic concepts

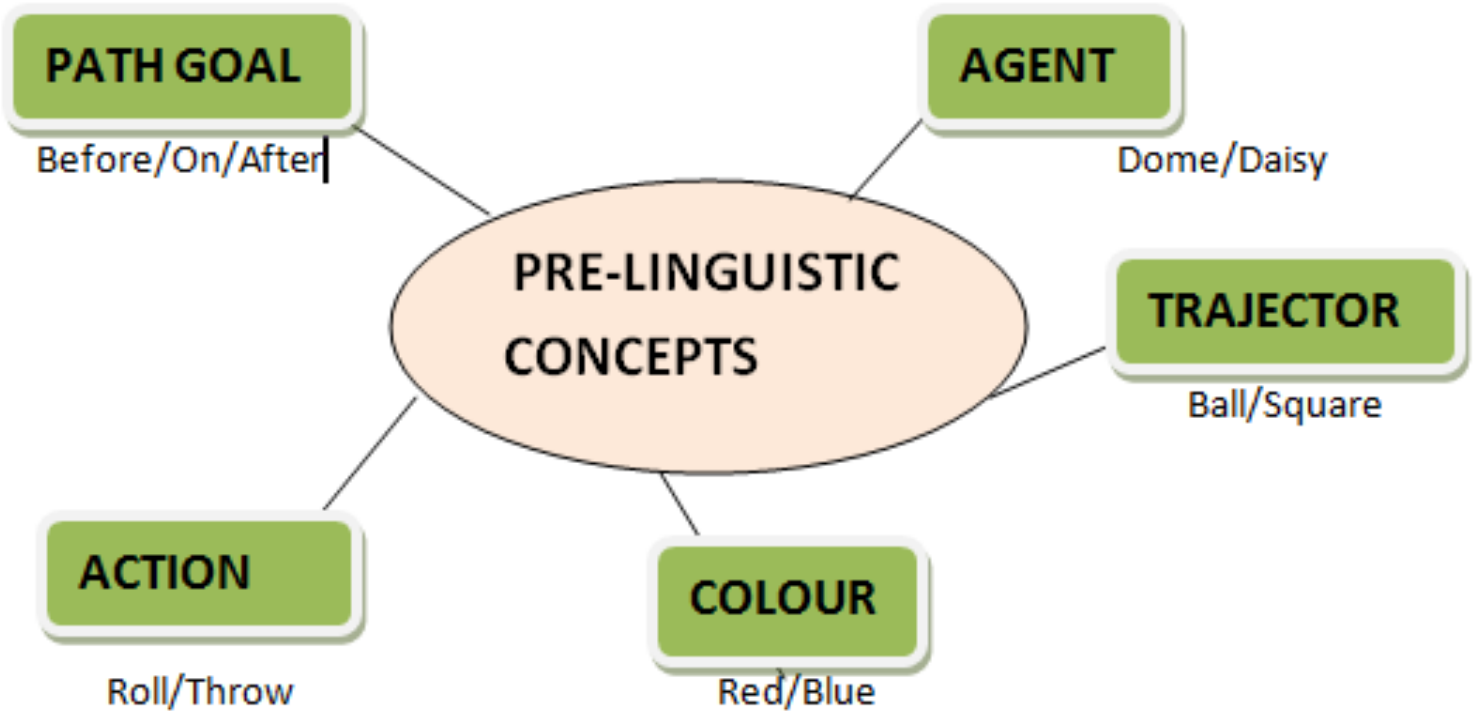
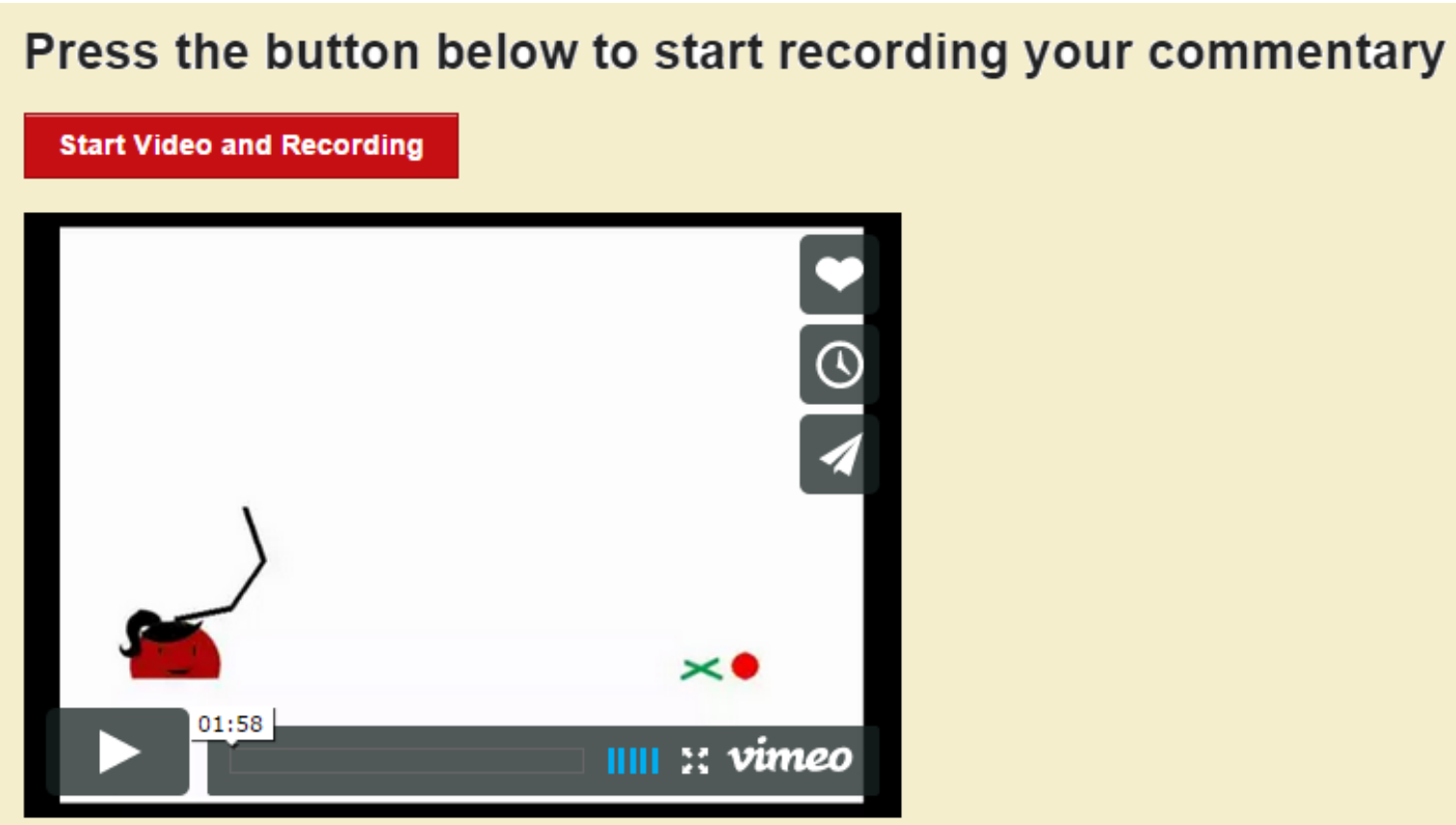


Figure : The pre-linguistic concepts

The interface

A snapshot of the interface for recording commentary in Bengali:



The commentaries

A sample of the Bengali commentaries collected:

Family-lect and Multi-lect

- Family-lect : the set of commentaries which are coherent with respect to the lexical choices for trajectors, actions, agents and the constructional choices.
- Multi-lect: this corpus includes the different varieties of syntax and vocabulary available.

Contrastive Association

For two non-overlapping concepts c_1, c_2 we first divide the corpus into subsets:

1. those that arise on commentaries for video involving c_1
2. those arising for c_2

Now a scoring function for association is defined as the ratio of the joint probabilities of word w occuring with concept c1 and that with c2:

$$S_{w,c_1} = \frac{P(w, c_1)}{P(w, c_2)}$$

ADIOS

An overview of ADIOS(Automatic Distillation Of Structure).
Algorithm:

- Initialization :loading all sentences
- for all i = 1 : N
  Pattern Distillation(i)
  Generalization(i)
  endfor
- repeat until no more significant patterns are found

Related Works

- Mukherjee et. al.[From visuo-motor to language, 2014] have shown how how a learning agent learns syntactic patterns based on some highconfidence words for English and Hindi.
- In D.Semwal's thesis [Dynamic NLP, 2014] we get an overview of the different dynamic NLP techniques that can be used.
- Z.Solan et. al. have outlined the ADIOS algorithm in[14].
- Zettlemoyer et. al. delve deep into morphosyntactic discovery in[14].

References

- D. Semwal, S. Gupta, A. Mukerjee. From visuo-motor to language (2014) At `http://www.cse.iitk.ac.in/users/amit/pub/aaai-fs_14.pdf`.
- D. Semwal. Dynamic NLP : A model for language bootstrapping and lifelong learning M.Tech. Thesis under Dr. Amitabha Mukerjee (2014). At `http://www.cse.iitk.ac.in/users/deepalis/thesis/report.pdf`.
- Z. Solan, D. Horn, E. Ruppin, S. Edelman. Unsupervised learning of natural languages. Proceedings of the National Academy of Sciences pp:11629-11634, 2005.
- A. Wang, T. Kwiatkowski, L. Zettlemoyer. Morpho-syntactic Lexical Generalization for CCG Semantic Parsing. In EMNLP, 2014.

CONCLUSION

QUESTIONS