Language Learning from Commentaries in Videos using Dynamic NLP

Anusha Chowdhury, Kaviti Sai Saurab
Guide: Prof. Amitabha Mukherjee

Department of Computer Science and Engineering, Indian Institute of Technology Kanpur

Problem

Learning a new language is tough even for a human being. Here we explore different techniques of Dynamic NLP through which an intelligent agent can learn a new language like Bengali or Telugu (word, pattern and syntax learning) from video commentaries.

Our Approach

- We collected 18 commentaries in Bengali and 9 in Telugu for a particular video.
- Next step was to construct a histogram for each data set and group them into clusters by applying the **Bhattacharyya distance**.
- The corpus is now divided into:
  - **Family-lect** : commentaries which have a lot of similarity
  - **Multi-lect** : the variations come under this category.
- Now perform **Contrastive Association** and system learns some high-confidence words for each concept.
- To **prune** the corpus, we cut down the words with high frequency and low contrastive score or extremely low freq.
- We use **word2Vec model** to learn the synonyms present in the corpus.
- For pattern and syntax learning, we applied the **ADIOS** algorithm.

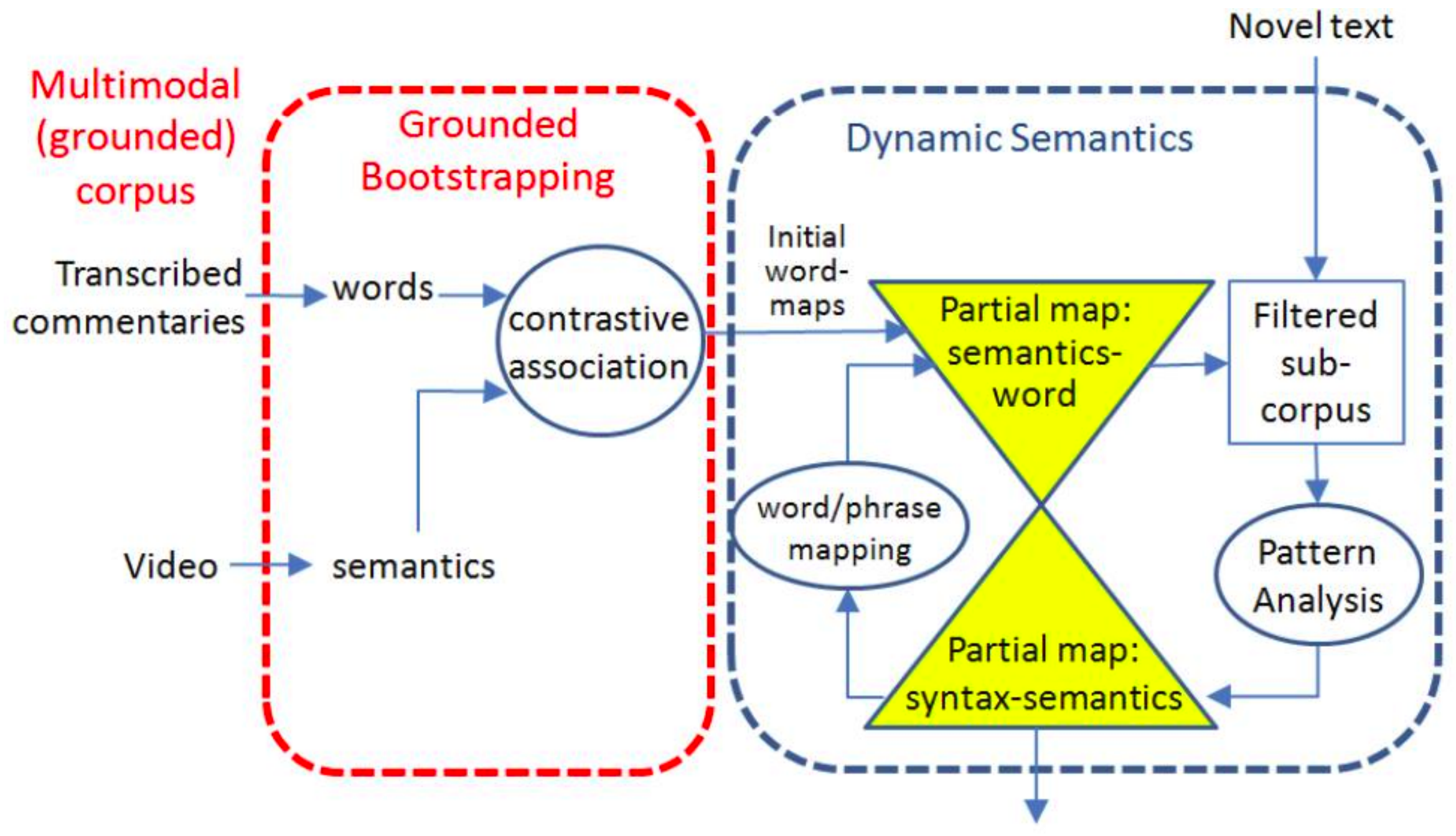


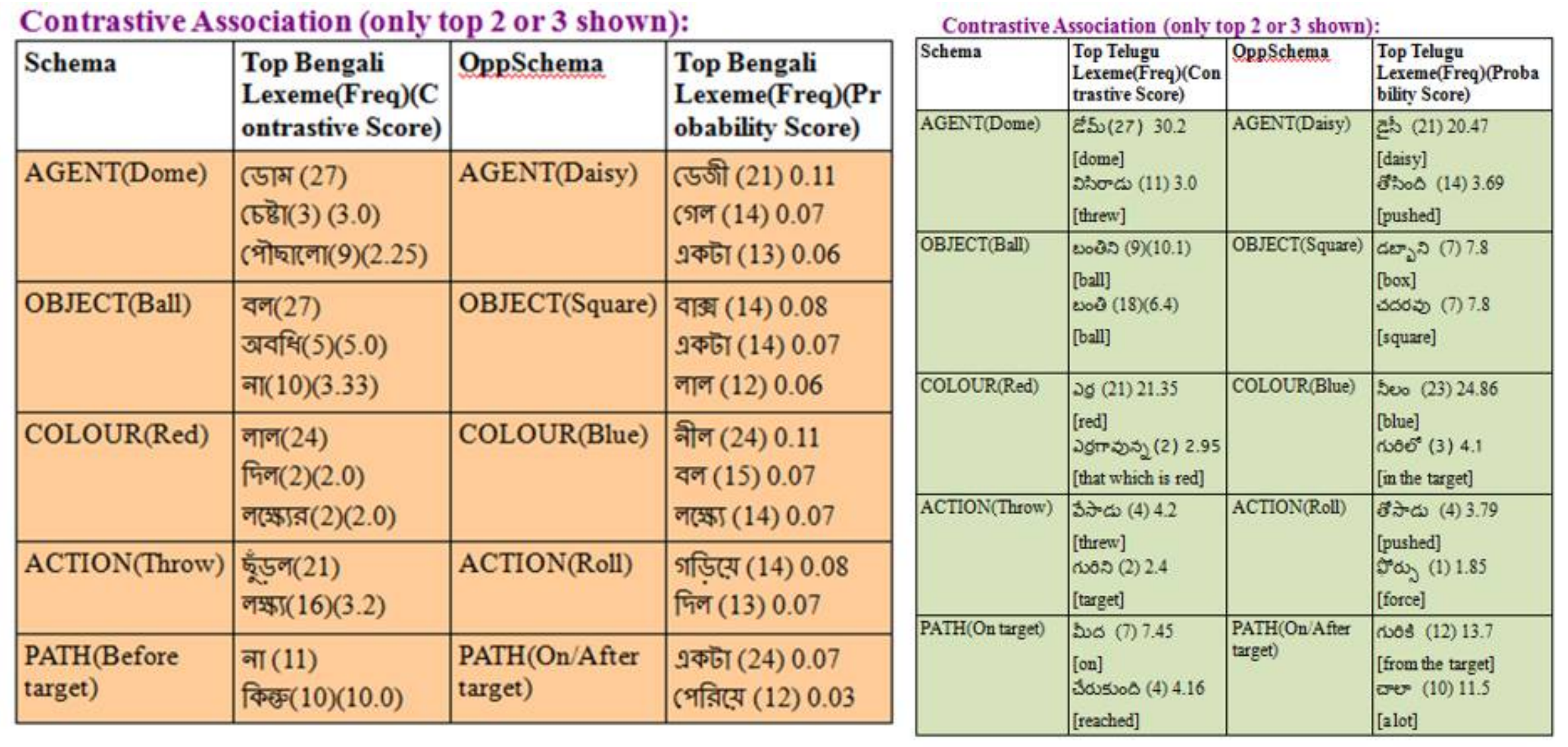
Figure 1: Dynamic NLP approach overview [1]

Contrastive Association

For disjoint concepts c_1, c_2 we divide the sentences into sets - those that occur in commentaries for video involving c_1 , and those that occur for c_2 [4].

$$Score_{w_i,c_1} = P(w_i, c_1)/P(w_i, c_2) \qquad (1)$$

The following results were obtained for Bengali:



Contrastive Association (only top 2 or 3 shown):

Schema	Top Bengali Lexeme(Freq)(Contrastive Score)	OppSchema	Top Bengali Lexeme(Freq)(Probability Score)
AGENT(Dome)	ডোম (27) চেষ্টা(3) (3.0) পৌছালো(9)(2.25)	AGENT(Daisy)	ডেজী (21) 0.11 গেল (14) 0.07 একটা (13) 0.06
OBJECT(Ball)	বল(27) অবধি(5)(5.0) না(10)(3.33)	OBJECT(Square)	বাক্স (14) 0.08 একটা (14) 0.07 লাল (12) 0.06
COLOUR(Red)	লাল(24) দিল(2)(2.0) লক্ষ্যের(2)(2.0)	COLOUR(Blue)	নীল (24) 0.11 বল (15) 0.07 লক্ষ্যে (14) 0.07
ACTION(Throw)	ছুঁড়ল(21) লক্ষ্য(16)(3.2)	ACTION(Roll)	গড়িয়ে (14) 0.08 দিল (13) 0.07
PATH(Before target)	না (11) কিন্তু(10)(10.0)	PATH(On/After target)	একটা (24) 0.07 পেরিয়ে (12) 0.03

Contrastive Association (only top 2 or 3 shown):

Schema	Top Telugu Lexeme(Freq)(Contrastive Score)	OppSchema	Top Telugu Lexeme(Freq)(Probability Score)
AGENT(Dome)	డోము(27) 30.2 [dome] విసిరాడు (11) 3.0 [threw]	AGENT(Daisy)	డైసీ (21) 20.47 [daisy] తోసింది (14) 3.69 [pushed]
OBJECT(Ball)	బంతిని (9)(10.1) [ball] బంతి (18)(6.4) [ball]	OBJECT(Square)	డబ్బాని (7) 7.8 [box] చదరపు (7) 7.8 [square]
COLOUR(Red)	ఎర్ర (21) 21.35 [red] ఎర్రగావున్న (2) 2.95 [that which is red]	COLOUR(Blue)	నీలం (23) 24.86 [blue] గురిలో (3) 4.1 [in the target]
ACTION(Throw)	వేసాడు (4) 4.2 [threw] గురిని (2) 2.4 [target]	ACTION(Roll)	తోసాడు (4) 3.79 [pushed] ఫోర్సు (1) 1.85 [force]
PATH(On target)	మీద (7) 7.45 [on] చేరుకుంది (4) 4.16 [reached]	PATH(On/After target)	గురికి (12) 13.7 [from the target] చాలా (10) 11.5 [a lot]

Figure 2: Contrastive Association on Bengali,Telugu

Histogram Similarity using Bhattacharyya distance

To measure the similarity of two discrete probability distributions, we calculate the Bhattacharyya coefficient as:

$$Bhattacoef = \sum_{i=1}^{n} \sqrt{(\sum a_i)(\sum b_i)} \qquad (2)$$

where samples a and b have n partitions, and $\sum a_i$, $\sum b_i$ are the no. of members of samples a and b in i^{th} partition. For discrete distribution $Bhattadistance = -ln(Bhattacoef)$.

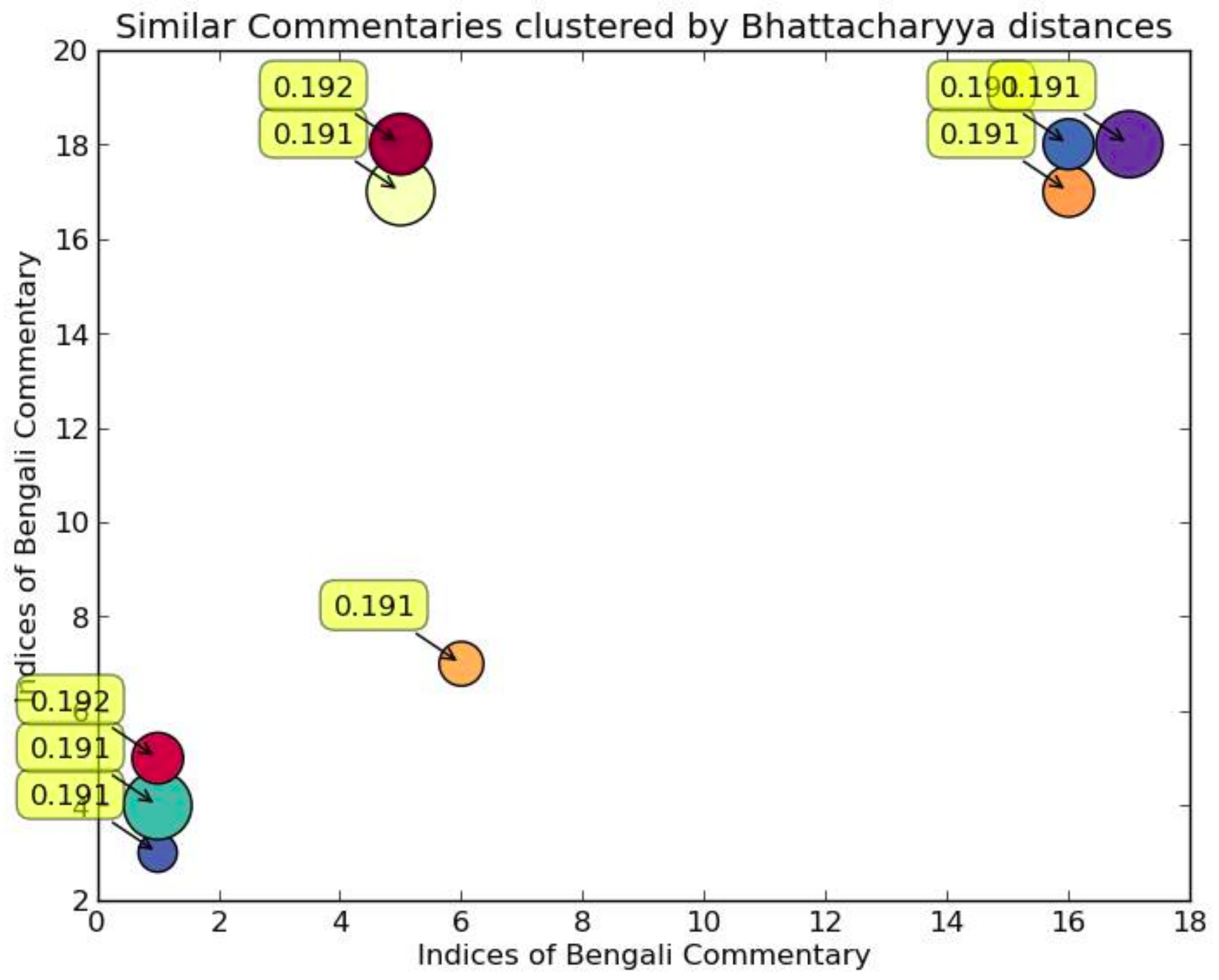


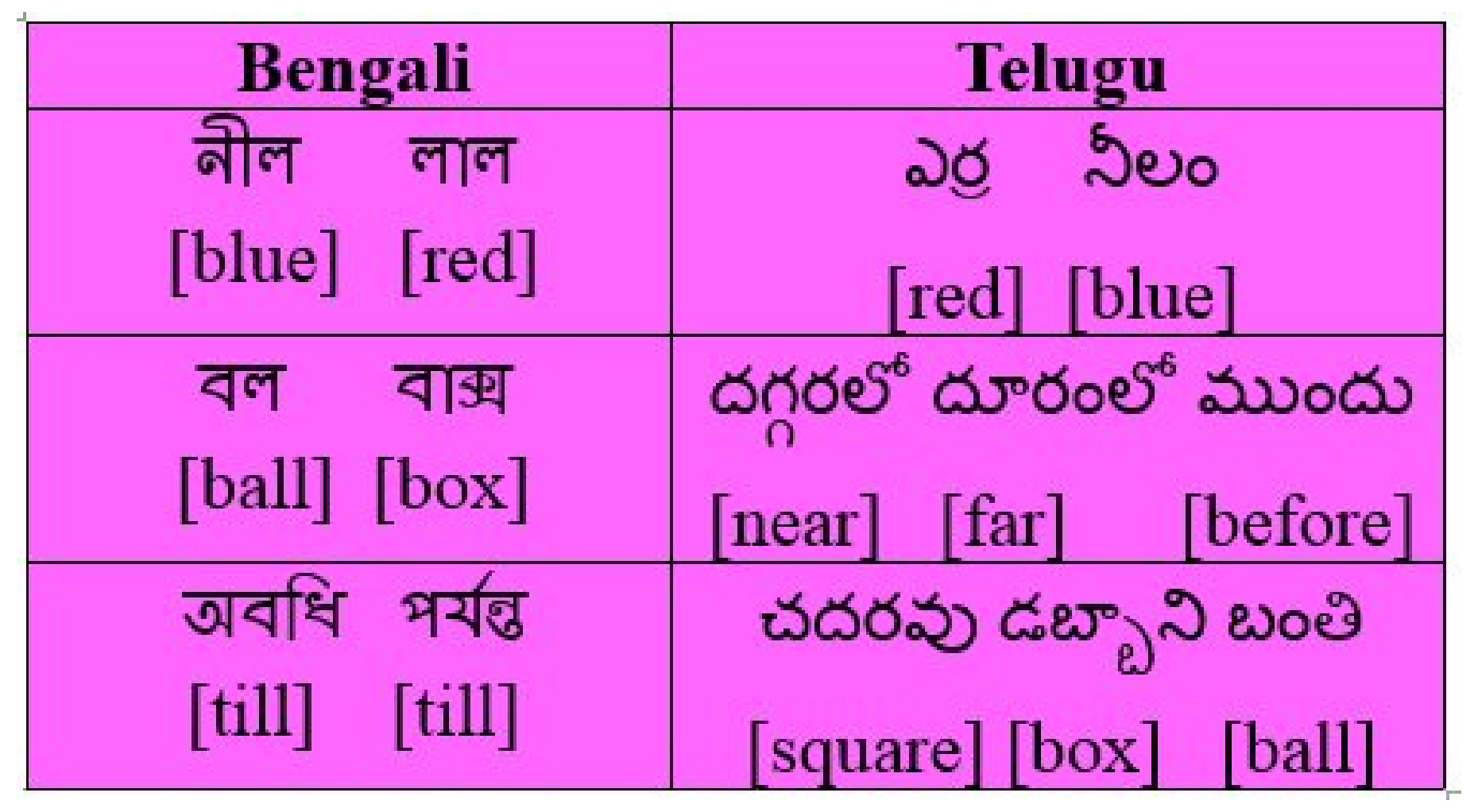
Figure 3: Clustering of Histograms from Bhattacharyya dist.

ADIOS(Automatic Distillation of Structure)

Adios works in an unsupervised fashion to extract significant patterns from a raw unannotated corpus. It works based on two principal components.

⋆ **Representational Data Structure**
⋆ **Pattern Acquisition Algorithm**

Probabilistic inference of pattern acquisition is used to cluster similar lexicons and recursive construction is used to hierarchically generate more complex patterns. Significance of patterns is determined using the **motif extraction criterion** [3]. At the end of each iteration, the most significant pattern is added to the lexicon as a new unit and the graph is rewired by merging the paths it subsumes. Algorithm stops when new patterns or equivalence classes are not discovered anymore.

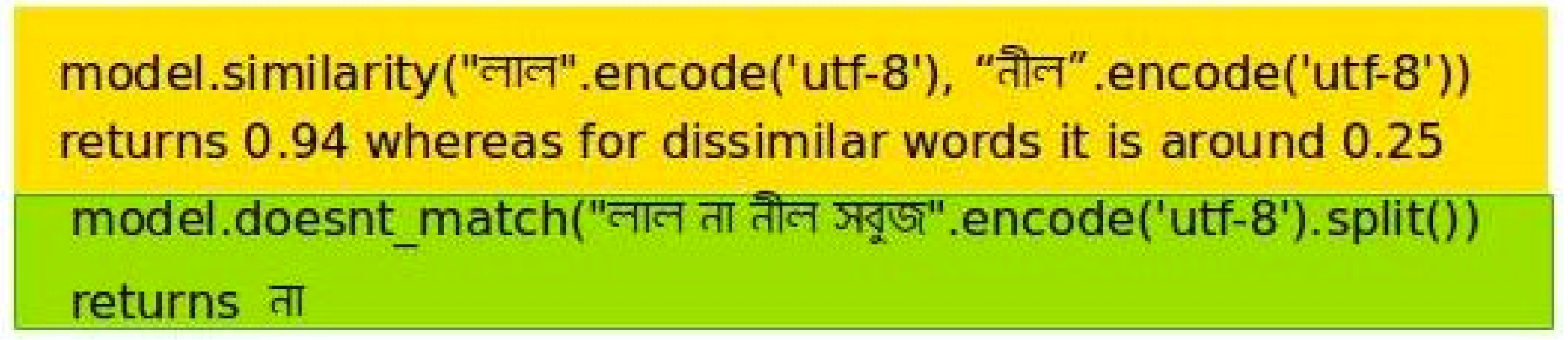


Bengali	Telugu
নীল লাল [blue] [red]	ఎర్ర నీలం [red] [blue]
বল বাক্স [ball] [box]	దగ్గరలో దూరంలో ముందు [near] [far] [before]
অবধি পর্যন্ত [till] [till]	చదరపు డబ్బాని బంతి [square] [box] [ball]

Figure 4: Equivalence classes obtained from ADIOS

Applying word2vec model on our corpus

The word2vec tool provides an efficient implementation of continuous **bag-of-words** and **skip-gram** architectures for computing vector representations of words. Here we apply it on our small corpus and interestingly we get quite good results in Bengali. It has learnt synonyms and also it can find the odd word out from a given list.



```
model.similarity("লাল".encode('utf-8'), “নীল”.encode('utf-8'))
returns 0.94 whereas for dissimilar words it is around 0.25
model.doesnt_match("লাল না নীল সবুজ".encode('utf-8').split())
returns  না
```

Figure 5: Results from word2vec

RESULTS OF ADIOS

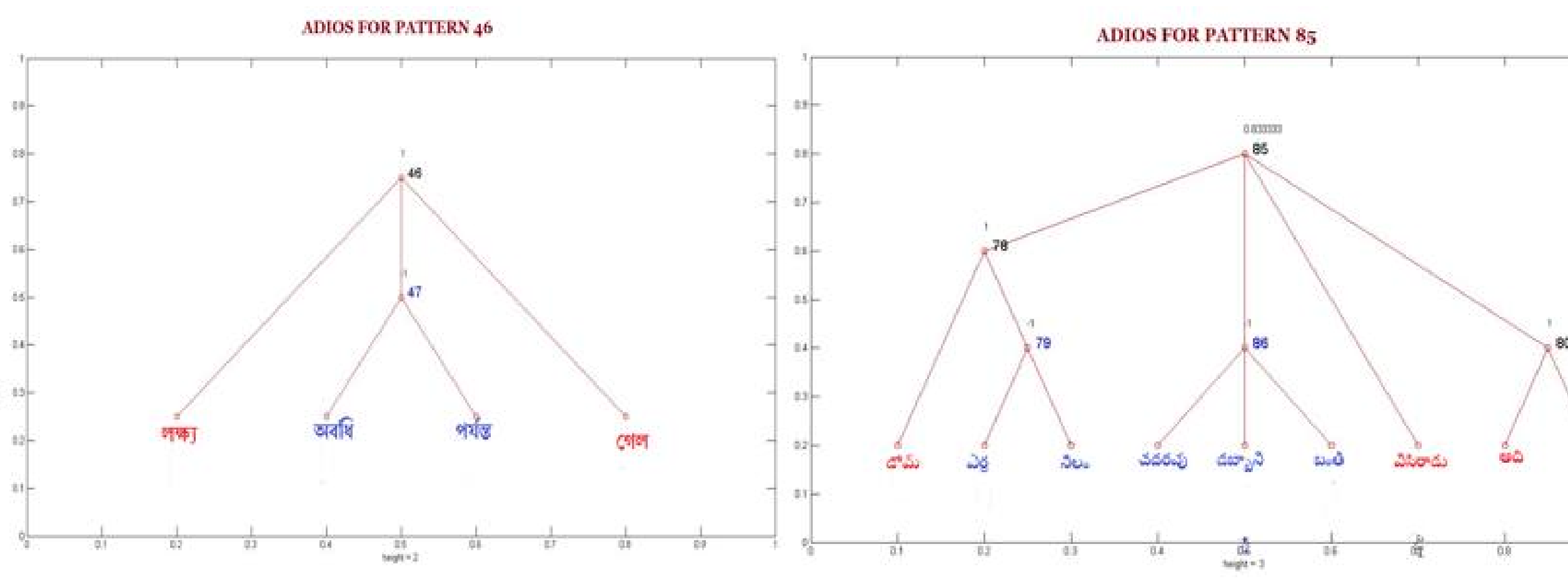


Figure 6: Results from ADIOS for Bengali,Telugu

Conclusion

Comparison with already existing work:

- [1] applies dynamic NLP on Hindi and English languages. Both [1] and [2], the grounding of target did not achieve high confidence. But Telugu,Bengali corpus yielded high confidence association of on,after target.
- Apart from ADIOS we also used word2Vec to study equivalence classes and performance was really good.

We are implementing **Morphosyntactic discovery** as final step where we will do a text-based morphological similarity analysis using Levenshtein distances between words (for example, in Bengali *lokhyo* and *lokhye* both refer to the same semantic concept target). These variations in pattern can be due to tense or gender difference in a language.

References

[1] D. Semwal. Dynamic NLP : A model for language bootstrapping and lifelong learning M.Tech. Thesis under Dr. A. Mukerjee at IIT Kanpur, 2014.

[2] D. Semwal, S. Gupta, A. Mukerjee. From visuo-motor to language.(2014)

[3] Z. Solan, D. Horn, E. Ruppin, S. Edelman. Unsupervised learning of natural languages(2005).

[4] E. M Markman. Constraints children place on word meanings.