CS365A - ARTIFICIAL INTELLIGENCE
Language Learning from Video Commentary in Bengali
Guide: Prof. Amitabha Mukherjee

Anusha Chowdhury(12148)

Indian Institute of Technology, Kanpur
March-April, 2015

Abstract. The aim of this project is to use a set of commentaries for the purpose of Bengali language learning. For word learning and syntax learning, a set of commentaries is collected on videos where there are agents, objects with colour, actions and path-goal. We use some techniques of **Dynamic NLP**(Natural Language Processing) including algorithms like ADIOS (automatic distillation of structure) to identify significant segments on the basis of statistical information.

Keywords: Dynamic NLP, ADIOS, Word learning, Syntax learning.

1 Introduction and Motivation

For a child, learning happens through language and through semantics. In dynamic NLP, we always associate a lexical discovery with its semantics. Dynamic refers to the fact that this method allows expansion of the obtained language model. Syntax deals with the positioning of the Subject, Verb and Object, which is generally in the order (Subject,Object,Verb) in Bengali. Being a Bengali myself, the motivation was to construct a model whereby the system will learn the words and syntax for this language. Understanding the concepts involved here would prove beneficial if we want to extend the project to any other language. The first step is to collect input data as mentioned in section 4. After that, to handle the variations in the input, a small subset is used for initial learning(Family-lect) and then the rest of the commentaries (Multi-lect) are considered. We apply ADIOS and other concepts like bigram, ngram, pruning of the corpus, etc.

2 Related Works

The ADIOS algorithm has been outlined in [1]. This unsupervised algorithm recursively distills hierarchically structured patterns from a given corpus. Susan et al. [5] introduce the concept of using transitional probabilities between words to segment sentences into phrases, and then use this information to acquire the syntax of a miniature language.
A language learning model has been established for English and Hindi using concepts of dynamic NLP [2]. The process has been demonstrated on a simple video, with crowdsourced commentary text in these two languages. Here we use similar concepts to design a language learning model for Bengali. In [4] we get a nice overview of the way a child picks up word meanings.

3 Our Approach

The learner has some pre-linguistic concepts in Bengali:

- Agent with subtypes ডোম, ডেজী (Dome,Daisy)
- Object with subtypes বল, বর্গক্ষেত্র (Ball,Square)
- Colour with subtypes লাল, নীল (Red,Blue)
- Action with subtypes ছোড়া, গড়িয়ে দেওয়া (Throw, Roll)
- Path-goal with subtypes লক্ষ্যর আগে, লক্ষ্য, লক্ষ্যর পরে (Before Target, On Target, Beyond Target)

Words like আর, এবং ('and') are not taken under consideration to narrow down the set of words. 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. This has been referred to as contrastive association in [2]. All these are on the family-lect stage. In the multi-lect stage, we prune the corpus and apply ADIOS algorithm.

4 Dataset

The input data consists of Bengali commentary for the sixteen 2D videos. Data is being collected from people of different genders and different ages in the campus. An example commentary from the collected samples in Bengali: ডেজী লাল বর্গক্ষেত্র গড়িয়ে দিল which can be translated as Daisy rolled red square. Till now, five sets of samples have been collected in Bengali and we will collect more as the project progresses.

References

1. Zach Solan, David Horn, Eytan Ruppin, Shimon Edelman. Unsupervised learning of natural languages. Proceedings of the National Academy of Sciences of the United States of America, 102(33):11629-11634, 2005.
2. Deepali Semwal. Dynamic NLP : A model for language bootstrapping and lifelong learning M.Tech. Thesis under Dr. Amitabha Mukerjee at IIT Kanpur, 2014. Available at `http://www.cse.iitk.ac.in/users/deepalis/thesis/report.pdf`.
3. Deepali Semwal, Sunakshi Gupta, Amitabha Mukerjee. From visuo-motor to language.(2014) Available at `http://www.cse.iitk.ac.in/users/amit/pub/aaai-fs_14.pdf`.
4. Ellen M Markman. Constraints children place on word meanings. Cognitive Science, 14(1):57–77, 1990. Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.9253&rep=rep1&type=pdf`.
5. Susan P Thompson, Elissa L Newport. Statistical learning of syntax: The role of transitional probability. Language Learning and Devvelopment, 3(1):1–42, 2007. Available at `http://www.ehu.eus/HEB/KEPA/teaching/NeuroAdvance2011/Thompson.Newport.statistical.learning.of.syntax.2007.pdf`.