Bayesian learner based Language Learnability Analysis of Hindi

Sandeep Saini*, Nitin Gupta*,
Shivin Bhogal†, Shubham Sharma†
*Department of Electronics and Communication Engineering
†Department of Computer Science and Engineering
The LNM Institute of Information Technology
Jaipur, India
(sandeep.saini,Y14UC173,Y12UC246,Y14UC287)@lnmiit.ac.in

Vineet Sahula
Department of Electronics and Communication Engineering
Malviya National Institute of Technology
Jaipur, India
sahula@ieee.org

***Abstract*—Language acquisition is one of the first tasks performed by the human brain. Linguistics has always been debating whether learning a native language is language dependent process or universal rules can be applied to learn all the languages. In this work, we analyse the learnability of Hindi based on Bayesian learner models. Considering the relation between theories of knowledge representation and theories of knowledge acquisition, Bayesian models are developed which can be modified to suit universal rules or language based rules. We verify the learnability of Hindi based on Unigram and Bigram models of learning. It is found that Hindi is better learnable using constrained learner models and shows better results in Bigram approach. We discuss implications for both theories of representation and theories of acquisition.**

***Keywords*—Language acquisition, Bayesian modeling, learnability analysis, Computational Intelligence**

I. Introduction

The human brain is having innate power to learn from the sounds and generate words [1]. These words, when produced with proper rules (grammar) make a language. When we talk specifically about the native language learning, then parents/guardians are the principal trainer for most of the infants. They speak words to get the attention of an infant or direct him/her to do some actions. How does infant brain decode these sounds and the brain processes these signals to trigger actions in different parts of the body? Which words/phrases are learnt faster/better by the infants? Can he/she learn all the words with equal easiness? Can any child learn any language? All these questions have been the topics of research in linguistics and recently computer scientists as well.

Berwick et.al [2], suggests that language serves as a cornerstone for human cognition, yet much about its evolution remains puzzling. Various language learning theories have been developed by linguistics to prove the learning models adopted by brain to learn/acquire any language [3] [4] and [5]. In this paper, we would focus on evaluating language learning strategies in Indian languages, specifically on Hindi.

Lisa Pearl et. al [6] have explored and evaluated the language learning strategies on 7 languages, namely English, German, Spanish, Italian, Farsi, Hungarian and Japanese. Liability of any language is measured in terms of F Score. F score is calculated from Precision and Recall capability of the learner model. Detailed description of F score is given in section V.

The paper is organized as follows. Section II provides an introduction to various language acquisition theories proposed by linguistics. Section III explores the computational models to evaluate these acquisition strategies. In section IV, we provide the details of data and preprocessing techniques used to filter the Hindi data for learnability analysis. We discuss the results in section V and conclude the paper in section VI.

II. Language Acquisition Theories

Hyams and Nina [7] states that "The ultimate issue in linguistic theory is the explanation of how a child can acquire any human language." An infant, coming from any social/economical background of the world, shows almost same capabilities to sit, grip, walk and talk. There are no serious differences in acquiring these skills in the initial stages of human life. While all other skills are identical across all infants, native language acquisition is different in all parts of the world. We focus on early stages of life in the process of language acquisition. Various stages of language learning at an early stages of the human life are as follows:

- From birth to 6 months of age. This stage is called prelinguistic stage and infants have almost no linguistic knowledge in this stage.
- From 6-8 months onset of babbling. In this stage infants shows the first manifestation of phonology.
- At around 10-12 months first words are generated by the infants. This stage is also critical to examine language generation process.
- At around 20-24 months onset of the two-word stage. Infants build vocabulary of native language and are able to interact in short answers.
- Till about 36-40 months: so called telegraphic speech. In this stage children have commendable grip over native language and good communication capabilities.

Macwhinney et. al [8] suggests that the Principle of contrast plays an important role in language acquisition in children. Principle of Contrast states that any difference in form in a

978-1-5090-2029-4/16/$31.00 @2016 IEEE

language marks a difference in meaning. This principle is a general one for speakers of a language. It is one that has been stated or assumed by virtually every linguist over the years. Based on the principle of contrast, Macwhinney have made 3 predictions.

- Children assume words contrast in meaning.
- Children give priority to known words.
- Children assign novel words that they hear to gaps in their lexicon, and, to fill such gaps, they coin new words themselves.

Anna Uhl Chamot [9], raises her concerns about issues in Language Learning Strategy Research and Teaching. In her extensive study, she compares three models for language learning strategy instruction. All three models SSBI Model [10], CALLA Model [11] and Grenfell & Harris model [12] are developed for students in early stages of their school life.

All three models involve identifying students' current learning strategies. These are identified through activities such as engaging in discussions about familiar tasks and completing questionnaires. Conclusions from these models suggested that the instructor should adopt new strategies for different languages. The CALLA model [11] is recursive in nature and teachers and students always have the option of revisiting prior instructional phases as needed. The Grenfell and Harris (1999) model [12], involves the students to work through a cycle of six steps, then begin a new cycle. The Cohen (1998) model has the teacher to take on a variety of roles in order to help students learn to use learning strategies appropriate to their own learning styles.

Recently people have started exploring Indian language learning strategies as well. Shagufta Khan [13] worked on utilizing mobile devices in Hindi/Urdu learning. He observed how self-organized learning environments (SOLE) and support universal design for learning (UDL) were differentiated by varying learning styles.

Most of these strategies are proposed and developed by linguistics or psychologists working closely with language learnability. With the evolution of computers, Natural language processing became one of the most researched topics [14] and [15]. Research groups are working toward Natural language programing. To achieve this language translation from natural language to machine language is a must. Therefor a lot of work is underway to map the human language learning strategies with machine learning strategies to develop the systems which would process the instructions in a way our brain in processing. In the next section, we provide the overview of some computational models and methods to evaluate language learning principles adopted by infant brain.

III. Evaluation Models

Language learning strategies mentioned in the above section are proposed and developed, with the assumption that they are universal in nature and would work in multiple languages. These strategies are mainly implied to human beings in some language learning centers the results are generated from data obtained from these evaluations. These evaluation results are dependent on human data and the whole process takes a lot of time and efforts in setting up the experiment and evaluation of the results.

In this section we explore some of the computational models and methods which can be used to evaluate language learning and acquisition capabilities of a person. We explore these language learning Strategies and how they work in different languages. Most of these strategies are tested on single language and very rarely cross-language analysis is performed. Demonstration of cross linguistic success of the strategies is very important in the infant stages of human life most of the theories presumed that the human being has innate properties to acquire and language as a newborn infant does we look at the computational models proposed for cross linguistic language learning strategies.

Most of the computational models to evaluate language learning strategies are Bayesian theory based. The complete process involves generating data from the Child-directed speech and get the inference from this data about the learnability of a particular language based on the proposed model. First, we'll discuss speech segmentation strategy and then the inference mechanism.

A. Bayesian Segmentation

We have investigated one version of the Bayesian segmentation strategy, that is, universal in nature and thus have language independent properties. Sound generated in any language is comprised of morphines, which are common for most of the languages. This model would be best suited for our target audience as well, which are infants in the age group of 6-12 months. At this age child is not having any prior knowledge of language and he learns from the sounds being produced around him. One benefit of Bayesian learning strategies is that they explicitly distinguish between the learners pre-existing beliefs (the prior: P(h)) and how the learner evaluates incoming data (the likelihood: $P(d|h)$). This information is combined using Bayes theorem (1) to generate the updated beliefs of the learner (the posterior: $P(h\|d)$).

$$P(h|d) \propto P(d|h) * P(h) \tag{1}$$

The Bayesian segmentation strategy was originally described by Goldwater et al. [16]. Considering the constraint that infants posses limited knowledge of language structure at the relevant age, GGJ described two simple generative models.

1) A unigram assumption
2) A Bigram assumption

Unigram model assumes independence between different words so effectively believes that word tokens are randomly selected. To encode this assumption in the model, GGJ assumes that the observed sequence of words $w_1...w_n$ is generated sequentially using a probabilistic generative process. To choose the identity of ith word in the Unigram case, Goldwater et. al uses equations (2) and (3), where the probability of the current word is a function of how often it has occurred previously.

$$P(w_i|w_1...w_{i-1}) = \frac{n_{i-1}(w_i) + \alpha * P_0 * (w_i)}{i - 1 + \alpha} \tag{2}$$

$$P_0 = P(w = x_i...x_m) = \prod_j P(x_j) \tag{3}$$

P_0 is a base distribution, specifying the probability that a novel word will consist of particular units $x_1...x_m$. $n_{i1}(w_i)$ is the number of times word w_i appears in the previous $i-1$ words, α is a free parameter of the model which encodes how likely a novel word is to be generated.

P_0 provides the model with a preference for shorter words. This can be understood from the fact that the total probability depends on the number of units (syllables) in a word. More units in a word would reduce the probability of choosing that particular word. Thus a word with smaller length will be the better learnt. Since children are more familiar and comfortable with smaller words, hence learnability is better. Therefore this method can be useful for language learning at earlier stages of human life.

Bigram model is slightly complex in its basic structure. It does not assume independence of consecutive words, but it assumes that consecutive words are related to each other and if they appear once, then there is a higher probability that they will appear again in the corpus. Thus a word is generated based on the identity of the word that immediately precedes it.

$$P(w_i|w_1 = w', w_1...w_{i-2}) = \frac{n_{i-1}(w', w_i) + \beta * P_1 * (w_i)}{n_{i-2}(w') + \beta} \quad (4)$$

$$P_i(w_i) = \frac{b_{i-1}(w_i) + \gamma * P_0 * (w_i)}{b - 1 + \gamma} \quad (5)$$

Here $n_{i-1}(w', w_i)$ defines the number of times a particular bigram (w_0, w_i) has occurred in the first i 1 words, $n_{i2}(w_0)$ provides the information on the number of times the word w_0 occurs in the first $i-2$ words, $b_{i1}(w_i)$ is the number of bigram types which contain w_i as the second word, b is the total number of bigram types which have existed before. P_0 was already defined in equation (3), and β and γ are free model parameters.

The role of $\beta and \gamma$ is very similar to the role of α in unigram model. They control the introduced bias towards fewer biagrams (β) towards fewer unique lexical items as the second word in a bigram (γ).

It is evident from the above equations and definitions that both Unigram and Bigram models prefer smaller lexicons because those words appears more frequently. A learner designed using any of these models would infer based on the data and model parameters, which words appears more in the corpus and showing better learnability for smaller words.

B. Bayesian inference

To infer the learnability of a language based on the data generated from the child-oriented speech, different types of learners are designed. We are exploring and utilizing Bayesian learners in our study. There are two categories of Bayesian learners adopted for this study, which are explained below [17].

Ideal learner processes data in a batch and thus it assumes that there is a perfect memory available for the learner. This learner has enough processing resources to exhaustively search potential segmentations. After an exhaustive search it selects optimal segmentation. This learner is heavily relying on the availability of huge corpus and works well if the learner have prior knowledge of the language. **BatchOpt** is the notation used for this ideal learner.

Constraint learner In practical world infinite memory (corpus) is not available and infants are also not having prior knowledge of the language. Thus a lot of constraints can be put on the learner as well. Three major types of constraint learners implemented in this work are:

1) Online Optimal: **OnlineOpt**
   This learner processes data incrementally. The learner have enough processing resources to exhaustively search potential segmentations, and it selects optimal segmentation.
2) Online Sub-optimal: **OnlineSubOpt**
   This constraint learner also processes data incrementally and have enough processing resources to exhaustively search potential segmentations. It differs from OnlineOpt learner in it's segmentation selection mechanism and it selects segmentation probabilistically.
3) Online Limited Working Memory: **OnlineMem**
   This learner also processes data incrementally like both constraint learners. But it has limited working memory buer, so it cannot do an exhaustive search. This learner focuses more on recent data (recency bias). It also selects optimal segmentation.

A summary of the properties of all ideal and constraint learners is complied in Table II.

multirow

TABLE I. SUMMARY OF BAYESIAN LEARNERS.

Learner	Parameters	Learning Assumptions		
		On-line processing	Sub-optimal Decisions	Recency Effect
BatchOpt	iterations = 20,000	No	No	No
OnlineOpt	N/A	Yes	No	No
OnlineSubOpt	N/A	Yes	Yes	No
OnlineMem	Sample uttrance = 20,000	Yes	No	Yes

IV. DATA PREPARATION FOR LEARNABILITY ANALYSIS

We have analyzed language learning models on Hindi. Corpus for Hindi is obtained from Center for Indian Language Technology (CFILT), IIT-Bombay [18]. This dataset contains Hindi literature related contents. A total of 1195 text files containing stories in Devanagari script. A total of 50,345 sentences are recorded in these files and total words are 345,675, which means that the average sentence length is 6.88 words. This data cannot be directly used in Bayesian learners. First, we have to transliterate the Devangari input into Roman script.

A. Transliteration

Transliteration is the practice of transcribing a word or text written in one writing system into another writing system. We needed to transliterate Devanagari words into Roman script because the further pre-processing steps have been developed to work for roman scripts. For transliteration algorithms used in [19] is involved. The algorithm has two major tasks to perform:

1) Provide single correct transliterated word in English script (Roman) if there is only a single word in Hindi script which could match with the original word.
2) Provide two-three most probable options of the transliterated text, in the order of higher to lower probability, if there is more than one words in Hindi script which could match with the original word.

Sample results obtained from the process of transliteration of Hindi text are shown in figure 1.

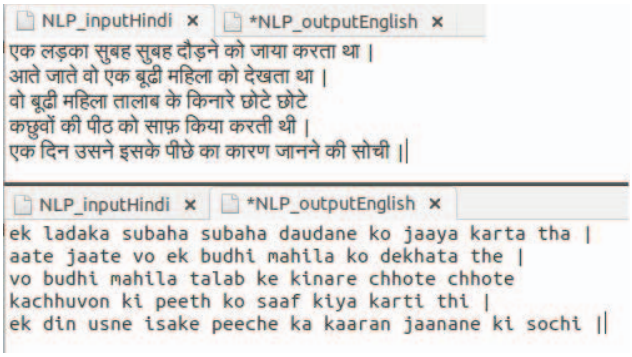

Fig. 1. a) Input Devanagari written Hindi sentences, b) transliterated Roman scripted sentences.

V. Results and Discussions

Language acquisition evaluation depends on many metrics proposed by different authors. These metrics should compare the adult orthographic representation, i.e., the way the words would appear when transcribed in the language by an adult speaker. In this process, we focus on how precisely the learner is able to recall the learnt words/syllables. Precision of learner is defined as:

$$Precision = \frac{identified\_true\_word\_tokens}{all\_identified\_word\_tokens} \quad (6)$$

While recall is the capability of the learner to recollect the learnt words, which is defined as:

$$Recall = \frac{identified\_true\_word\_tokens}{all\_true\_word\_tokens} \quad (7)$$

These two scores, which range between 0 and 1, are typically combined into a single summary statistic via the harmonic mean, referred to as the F-score.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

F score would also range between 0 and 1. This can be multiplied by 100 to get the evaluation score in percentage. Higher the value better would be the learnability of the learner model for a particular language. This metric is very simple to compute and assuming an orthographic transcript of the data is available, and so is a convenient way to compare the results of modeled segmentation strategies across studies. Still, a known disadvantage is that the target segmentation is assumed to be the adult orthographic segmentation, which is unlikely to be true for six- to seven-month-olds using early segmentation strategies.

F score calculations for English and Hindi languages are shown in Table II. These scores show that English is showing best learnability for Bigram model used by constraint learner. Thus English can be better learnt if the learner has more recent knowledge and next words are more dependent on the previous words. Hindi on the other hand, provides best F score for Bigram model combined with BatchOpt Bayesian learner. This implies that prior knowledge of the language and relation between consecutive words increases the learnability of the languages.

TABLE II. F score calculated for English and Hindi languages based on Unigram and Bigram based ideal and constraint learners.

Model	Learner	F Score	
		English	Hindi
Unigram	BatchOpt	0.531	0.459
	OnlineOpt	0.588	0.586
	OnlineSubOpt	0.637	0.589
	OnlineMem	0.551	0.605
Bigram	BatchOpt	0.771	**0.814**
	OnlineOpt	0.751	0.759
	OnlineSubOpt	0.778	0.785
	OnlineMem	**0.863**	0.741

VI. Conclusion and Future Work

In this paper, we have studied the learnability of Hindi based on language learning models. Bayesian models along with different n-gram models are employed to calculate the learnability metrics for the languages. It is concluded that unlike English, Hindi is best learnable with the ideal learner approach. This behaviour of Hindi is because of different grammar structure than English. The language acquisition is better if the user or learner has prior knowledge of the language. The results showed similarity with English in terms of Bi-gram approach. Both the languages are better learnable with Bi-gram models.

References

[1] Duane M Rumbaugh. *Language learning by a chimpanzee: The Lana project*. Academic Press, 2014.

[2] Robert C Berwick, Angela D Friederici, Noam Chomsky, and Johan J Bolhuis. Evolution, brain, and the nature of language. *Trends in cognitive sciences*, 17(2):89–98, 2013.

[3] Nelson Brooks. Language and language learning, theory and practice. 1964.

[4] H Douglas Brown and . Principles of language learning and teaching. 2000.

[5] Stephen D Krashen and Tracy D Terrell. The natural approach: Language acquisition in the classroom. 1983.

[6] Lisa Pearl. Evaluating learning-strategy components: Being fair (commentary on ambridge, pine, and lieven). *Language*, 90(3):e107–e114, 2014.

[7] Nina Hyams. *Language acquisition and the theory of parameters*, volume 3. Springer Science & Business Media, 2012.

[8] Brian MacWhinney. *Mechanisms of Language Acquisition: The 20th Annual Carnegie Mellon Symposium on Cognition*. Psychology Press, 2014.

[9] Anna Uhl Chamot. Issues in language learning strategy research and teaching. *Electronic journal of foreign language teaching*, 1(1):14–26, 2004.

[10] Andrew D Cohen. *Strategies in learning and using a second language*. Routledge, 2014.

[11] AU Chamot. The cognitive academic language learning approach (calla): An update. *Academic success for English language learners: Strategies for K-12 mainstream teachers*, pages 87–101, 2005.

[12] Michael Grenfell and Vee Harris. *Modern languages and learning strategies: In theory and practice*. Psychology Press, 1999.

[13] Shagufta Khan. Utilizing mobile computer devices in urdu/hindi language programs to enhance language learning. 2015.

[14] Sandeep Saini and Vineet Sahula. A survey of machine translation techniques and systems for indian languages. In *Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on*, pages 676–681. IEEE, 2015.

[15] Sandeep Saini, Umang Sehgal, and Vineet Sahula. Relative clause based text simplification for improved english to hindi translation. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pages 1479–1484. IEEE, 2015.

[16] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009.

[17] Lawrence Phillips and Lisa Pearl. Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the Computational and Cognitive Models of Language Acquisition and Language Processing Workshop*, pages 9–13, 2014.

[18] IIT-Bombay Hindi Corpus. http://www.cfilt.iitb.ac.in/downloads.html. 2010.

[19] Kanika Gupta, Monojit Choudhury, and Kalika Bali. Mining hindi-english transliteration pairs from online hindi lyrics. In *LREC*, pages 2459–2465, 2012.