

Syllable based noun recognition for grounded videos

Prashant Jalan

Department of Computer Science & Engineering

Indian Institute of Technology Kanpur

Homepage: www.prashantjalan.com

April 3, 2013

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 1 |
| 1.1 | Language learning framework | 1 |
| 1.2 | Resources exploited | 1 |
| 2 | Collecting the dataset | 2 |
| 2.1 | Collecting Narrations | 2 |
| 2.2 | Audio to text transcription | 3 |
| 3 | Learning the language labels | 4 |
| 3.1 | Attention Model | 4 |
| 3.2 | Linguistic segmentation | 4 |
| 3.3 | Association Measure | 5 |
| 3.3.1 | Joint Probability | 5 |
| 3.3.2 | Conditional Probability | 5 |
| 3.3.3 | Mutual Information | 5 |
| 3.3.4 | Relative Frequency | 5 |
| 4 | Corpus Analysis | 6 |
| 4.1 | Most frequent words in Hindi | 6 |
| 4.2 | Top k-grams in Hindi | 7 |
| 5 | Results | 8 |
| 5.1 | Hindi | 8 |
| 5.2 | English | 8 |
| 6 | Conclusion & Future Work | 9 |
| 6.1 | Conclusion | 9 |
| 6.2 | Future Work | 9 |

Acknowledgements

I would like to thank Dr. Amitabha Mukerjee under whose supervision I did this work. If not for his guidance, motivation and support, this work would not have been possible. I am thankful that he gave me an opportunity to work under him.

I am thankful to Mr. Diwakar Chauhan who was always there to help me understand the difficult concepts and motivate me throughout the project.

I acknowledge the excellent work of Nikhil Joshi in 'unsupervised language learning for complex 3D videos'.

I owe to my parents and my sister for their moral support and encouragement.

Finally, I thank IIT Kanpur for giving me this platform to pursue my interests and do this research work.

Abstract

We aim to make the computers learn a new language without any previous knowledge about the language. In this work, we have used a semantic syllabic approach and also a word level analysis to acquire basic linguistic units particularly, noun based on the Langacker [1] theory of learning language. Based on a 2D video and co-occurring raw text, we demonstrate how this cognitively inspired model segments the world to obtain a meaning space, and combines words into hierarchical patterns for a linguistic pattern space. We try to recognize nouns in the English language and the Hindi language based on some narrations taken from different subjects using different association measures such as the mutual information, relative frequency and conditional probability.

Chapter 1

Introduction

1.1 Language learning framework

The problem of language acquisition has been of great interest to many disciplines including Linguistics, Psychology, Philosophy, Neurobiology, Cognitive science and Computer Science. From Panini [1] to Chomsky [2] to Tomasello, there have been many attempts to formalize the theory of language. The debate is mostly two-sided. Chomsky [2] argues for the innateness of language based on the argument (known as ‘poverty of stimulus’) that the child acquiring language has access to only positive examples (grammatical sentences), and very little corrective feedback. Thus, the Chomskyan framework focuses on the syntax of a language and is largely sceptical about semantics. So, learning a language from his viewpoint is learning a ‘generative syntax’ for that language. Langacker [3], alternatively has given a central role to semantics in his language learning model. Langacker [3] considers grammar as conceptualization and formalizes it as a bipolar symbolic unit interconnecting the phonological pole (linguistic representation) and the semantic pole (conceptual representation). In the view of cognitive grammar, language is entrenched in the usage and linguistic representations get their meanings because of their usage with some conceptual entity. The idea is analogous to a child’s way of learning. When a child is born, he knows nothing about a language. He doesn’t know anything about the noun, verb, preposition or the syntax or the word boundaries. But as he continuously hears description, slowly after many instances of a particular object or an action being referred to by a particular word, the child begins to recognize the word and associate it with the object or action.

1.2 Resources exploited

We used python as the programming language because of its ability to handle large datasets and execute complex algorithms with very simple commands. We used Windows Movie Maker to record narration for the video clips and export it without any change in the video properties such as the frame number, bit rate and frame size. We had also used the VideoPad Video Editor software to embed the audio clip (recorded using the Windows default sound recorder software) onto the video clip. We used Microsoft Word or Open Office to manually transcribe the audio into written Hindi text. To write in Hindi we also used the online editor available at www.quillpad.in

Chapter 2

Collecting the dataset

We collected our dataset by asking different subjects from different linguistic backgrounds to describe the objects and their actions as they saw in the 2-D grounded video.

2.1 Collecting Narrations

We collected narrations in Hindi from twenty one different subjects. Out of them ten subjects had the words लाल and नीला in their narrations and eleven subjects had the words छोटा and बड़ा in their narrations. For English, we collected narrations from thirteen subjects. Ten subjects used the words **big** and **small** and five used **red** and **blue** in their narrations.

Each subject was given the same set of instruction, the instruction set being ‘You will be shown this 39 seconds video thrice. For the first two times you can just see the video and gain an understanding of what is happening in the scene. The third time you have to describe whatever is going on in the video in Hindi/English without involving yourself in the description. You are also not allowed to metaphorize the objects in the video.’

The subject couldnt involve themselves in the narration as then the narration would not be just describing two simple objects and their actions in the grounded video, rather it would become a more complex set of narration involving the objects and their actions in the video, the narrator and his actions during the narration.

Another crucial phenomenon we observed while taking narrations was the occurrence of an inadvertent time lag between the action going on in the video and the words describing it. This time lag varied from person to person and could be positive. The purpose of showing the video thrice was to reduce the time lag.

The speaker were also asked not to metaphorize the objects in the video when we found some speakers replacing the object with some other noun, for instance छोटा त्रिभुज with चोर and बड़ा त्रिभुज with पुलिस.

| | | |
|-----|-----|--|
| 1 | 13 | |
| 14 | 46 | एक चतुर्भुज में दो त्रिभुज हैं |
| 47 | 55 | |
| 56 | 72 | एक त्रिभुज लाल है |
| 73 | 102 | एक त्रिभुज नीले रंग का है |
| 103 | 160 | दोनों त्रिभुज एक दूसरे से लड़ने की कोशिश कर रहे हैं |
| 161 | 201 | और लाल त्रिभुज बाहर भाग गया है |
| 202 | 249 | फिर चतुर्भुज में आने का प्रयास कर रहा है |
| 250 | 311 | नीला त्रिभुज और लाल त्रिभुज एक दूसरे से टकरा रहे हैं |
| 312 | 388 | लाल त्रिभुज नीले त्रिभुज को बाहर धकेल रहा है |
| 389 | 466 | नीला त्रिभुज लाल त्रिभुज से लड़ने के लिए तैयार हो रहा है |
| 467 | 500 | नीला त्रिभुज बाहर बाहर ही घूम रहा है |
| 501 | 552 | लाल त्रिभुज उसके पीछे भागने की कोशिश कर रहा है |
| 553 | 598 | दोनों एक दूसरे से टकराकर घूम रहे हैं |

Figure 2.1: A narration in Hindi.

| | | |
|-----|-----|--|
| 1 | 18 | |
| 19 | 50 | these are basically two triangles |
| 51 | 67 | which are in a boundary |
| 68 | 117 | seems like the bigger red triangle is trying to push |
| 118 | 150 | the smaller blue triangle out of the boundary |
| 151 | 163 | |
| 164 | 200 | right now the red triangle is outside the boundary itself |
| 201 | 240 | and it is trying to do something |
| 241 | 280 | so that it can push the blue triangle outside the boundary |
| 281 | 353 | rather it has come behind the blue triangle |
| 354 | 402 | and it is pushing it back through the boundary |
| 403 | 469 | so it is quite successful but now it has blocked its path |
| 470 | 510 | so that it cannot re enter |
| 511 | 573 | the blue triangle is around so that it can |
| 574 | 598 | |

Figure 2.2: A narration in English.

2.2 Audio to text transcription

Since, the audio speech recognition for Hindi is not yet very efficient and as our primary aim is learning language, we manually transcribed the audio clips into written Hindi texts. While transcribing the utterances, every two consecutive words were separated by space to maintain word boundary. The post-positions were generally treated as separate words and hence were separated from the content words they were attached to. However, the morphological variations were preserved and transcribed as it is without separating them from their roots. While transcribing small grammatical corrections were made. The care was taken to follow uniform writing style to avoid the transliteration variations. We embedded the narration with a reference to the frame number and broke it whenever there was a pause or any sound not describing a proper Hindi word (breathing sound or any other sound due to an external stimulus). If any pause was more than five frames, we made a separate row stating the frame period and entered empty text (nothing). For less than five frames we would merge the pause in the starting of the next narrative sentence as it was not possible manually to achieve an accuracy of less than five frames pause.

Chapter 3

Learning the language labels

This chapter describes the theoretical concepts behind the noun recognition.

3.1 Attention Model

We use an attention model to find the most salient part of the scene. Such a model tries to predict the part of the scene the human is most likely to attend to. The words used in the description are more likely to refer to objects that are in perceptual focus, i.e. we assume that linguistic focus follows perceptual focus.

Our attention model is based on the findings that objects that are moving are likely to be more salient. We ignore some other factors such as color and texture, which are more relevant in still images; for image sequences, motion and size are more significant. Size and speed are also not a major factor in our attention model as the objects are not always moving together. For appreciable amount of time in the video, we find that when one object is moving the other isn't and therefore, the perpetual focus will mostly be guided by the moving object.

3.2 Linguistic segmentation

Linguistic segmentation refers to breaking down the utterances into smaller linguistic units. However, what the smaller linguistic unit of break-up should be is a debatable issue. In our syllabic analysis, we broke the text into syllables without any knowledge of the word boundaries and in our word analysis, we took a single word as one linguistic unit.

A unit of pronunciation having one vowel sound, with or without surrounding consonants, forming the whole or a part of a word is defined as a syllable. We have used the following FSM (Finite State Machine) to formulate an algorithm to find the syllabic units from the text assuming that we don't know the word boundaries.

In the FSM, {C}, {V}, {M}, {O}, {P}, {S}, {H}, {D} denotes the set of all consonants (क, ख, ग,...), vowels (अ, आ, इ, ई, ...), matras (का, की, ...), accompanying consonants in Hindi, the character अं, punctuation marks (।, ॥, ...), signs (', :, , ...), halant and the digits in the Hindi language respectively. The program starts from the *initial* state. In the *store* state whatever has been pulled out from the text list is stored as a syllable

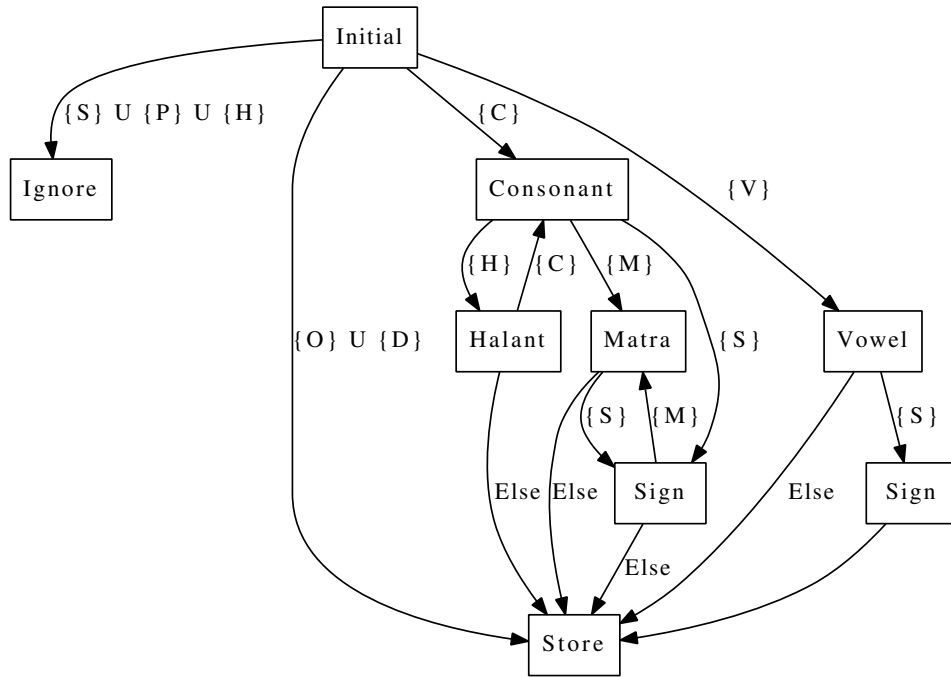


Figure 3.1: *FSM to identify the syllables.*

and in the *ignore* state it is ignored. From the *initial* state the machine goes to the *ignore*, *store*, *consonant* or *vowel* state depending on what it scans out of the list of text. Similarly, it proceeds forwards from the *consonant* and *vowel* state.

3.3 Association Measure

3.3.1 Joint Probability

3.3.2 Conditional Probability

3.3.3 Mutual Information

3.3.4 Relative Frequency

Chapter 4

Corpus Analysis

While learning the linguistic units, we did not consider the most common and frequent words used in English and Hindi. For English, we took the most commonly used words from a previously done analysis using British English Corpus, American English Corpus and recorded talks and speech [1]. For Hindi, we use Hindi unicode corpus, Center For Indian Language Technology, IIT Bombay [2]. We perform both syllabic and word analysis in the Hindi corpus to discover the most frequent words and top k-grams in Hindi.

4.1 Most frequent words in Hindi

The most frequently used words in Hindi are as follows:

| | | | | | | | | | |
|----|-----|----|------|----|------|----|------|----|------|
| 1 | के | 11 | भी | 21 | इस | 31 | थे | 41 | जा |
| 2 | है | 12 | नहीं | 22 | लिए | 32 | थी | 42 | रहा |
| 3 | में | 13 | कि | 23 | कर | 33 | न | 43 | मैं |
| 4 | की | 14 | एक | 24 | वह | 34 | कुछ | 44 | कोई |
| 5 | से | 15 | ही | 25 | किया | 35 | जाता | 45 | वे |
| 6 | और | 16 | हो | 26 | गया | 36 | साथ | 46 | हुए |
| 7 | का | 17 | तो | 27 | तथा | 37 | या | 47 | रूप |
| 8 | को | 18 | यह | 28 | करने | 38 | तक | 48 | किसी |
| 9 | हैं | 19 | था | 29 | जो | 39 | होता | 49 | हुआ |
| 10 | पर | 20 | ने | 30 | अपने | 40 | दिया | 50 | उसे |

4.2 Top k-grams in Hindi

The top 2,3,4,5,6 k-grams in Hindi are as follows:

| 2-gram | 3-gram | 4-gram | 5-gram | 6-gram |
|--|--|---|--|---|
| कर और पर इस ताहै एक नहीं उस लिए कार | केलिए करने अपने नेकेलिए जाताहै हैऔर यागया हैइस तकर उसके | नेकेलिए करनेके सकताहै कियागया आवश्यक सरकार केकारण रनेकेलि सप्रकार याजाताहै | करनेकेलिए रनेकेलिए इसप्रकार केअनुसार आवश्यकता जासकताहै कियागयाहै याजासकता कियाजाताहै आधारपर | करनेकेलिए याजासकताहै केआधारपर कीआवश्यकता कियाजासकता अलगलग उत्तरप्रदेश होनेकेकारण करोड़रुपये तकरनेकेलिए |

Chapter 5

Results

5.1 Hindi

5.2 English

Chapter 6

Conclusion & Future Work

6.1 Conclusion

Given the object categories discovered and visual saliency of these objects over the time, we demonstrate the ability of our system to learn nouns like त्रिभुज, त्रिकोण, लाल, नीला, बड़ा, छोटा in Hindi and triangle, red, blue, big, small in English. We confirm the success in learning words by analysing the strength of associations with increasing number of narrations. Discovering लाल and नीला from narrations describing the triangles as लाल and नीला and छोटा and बड़ा from the narrations which describe the triangles as छोटा or बड़ा, both in English and Hindi, confirms the success of our model. We argue that the consistent dominance of association strength of label with a visual category over the other labels is desirable and can be taken as a confirmation of the word learning. The success in learning appropriate labels even without knowing word-boundaries shows that the knowledge of word boundaries may not be a prerequisite for early word-learning. Getting the same results at a word level analysis illustrates the correctness of the association measures we have used. The results show that Hindi is a highly inflected language.

6.2 Future Work

We aim to extend our word to discover other linguistic units such as verbs, prepositions and finally, be able to learn a language with its syntactical knowledge. To enhance the results of noun discovering we wish to apply morphology to remove the word inflections. We hope to do a syllabic analysis for English, too. We also aim to use other association measures such as Dominance Weighted Joint Probability, which is proposed by Guha and is described in [1]. We also hope to get better results in English after collecting some more narrations.

Bibliography

- [1] <http://www.world-english.org/english500>
- [2] <http://www.cfilt.iitb.ac.in>
- [3] http://home.iitk.ac.in/~prasant/Corpus_files