

## 8.16 Rule-based Machine Translation System using Indian Logic for Discourse Texts

Kommaluri Vijayanand

**Abstract**—Communication is said to be taken place when the message delivered is received effectively at the receivers end. The communication channel and mode of communication are least significant than delivering the message to the targeted audience. Several texts appear in day to day life which carries messages to its targeted audience in the form of advertisements, billboards, posters, banners etc. Though the message delivered is precise in its form, such messages are received by the public effectively. No linguistic information is found in such texts. The mechanism applied by the people to understand such texts is the ‘Reasoning’ and ‘Interpretation’. Representing such reasoning and interpretation capabilities to the machine is one of the challenges to Artificial Intelligence Researcher. Among Indian philosophers who worked on Logic, Acharya Dignaga is a significant logician, also called a ‘Fighting Bull’. The present paper is aimed to present the issues in Critical Discourse Analysis (CDA) and Knowledge representation using Dignagas Nyaya-praveśa, towards Machine Learning along with the implementation details.

**Index Terms**—Word Sense Disambiguation, Indian Logic, Discourse Texts, Machine Learning, Reasoning, Knowledge Representation, Critical Discourse Analysis, Perception, Inference.

### I. INTRODUCTION

ARTIFICIAL Intelligence (AI) is a dynamic area of research, that provide scope for transforming an idiot (the Computer) into intelligent. In doing so, a computer scientist need to equip the computer with knowledge that is laid down in the society in various forms. Till now, several Machine Translation (MT) systems were reported in India and abroad using various approaches and techniques. These systems are fed with knowledge in the form of electronic dictionaries, the example-base with bi-lingual, tri-lingual and multi-lingual parallel corpus, the rule-base, etc. Question-Answering Systems, Interactive dialogue systems, Online Hotel and Railway reservation systems, etc. were also reported which rely upon AI that is fed in different forms. Most of the input, received in the form of text or speech from the general public to the system, does not meet the linguistic rules and requirements. The MT systems are prone to generate erroneous output due to ambiguity, when fed with such inputs. Under such circumstances, an urgent need to equip the computer to think on its own is identified. This is possible when the computer can judge based on logic. Towards such initiation, an attempt is made in this paper to identify the possibility of knowledge representation using Indian Logic towards handling the discourse texts.

Logic is delivered by several Indian philosophers and Logicians, dated back from 1500 B.C. in the form of Vedas. Among such great philosophers, Āchārya Dignāga is a sig-

nificant and earliest Buddhist writer on Logic, also called a “Fighting Bull” or a “Bull in discussion”. He was regarded as a father of Mediaeval Logic. The branch of learning was paid attention differentiated from general philosophy began in 450 A.D. Dignāga, in his writing, divided valid knowledge into six components, viz. Perception, Inference for ones’ own self, Inference for the sake of others, Reason and example, Negation of the opposite, and Analogue. Dignāga stated two pramāṇās (means of valid knowledge), namely, the perception and inference. These components of knowledge are further elaborated in section I-A.

The development of precise frameworks of discourse interpretation has been hampered due to the lack of deeper understanding of the dependencies between different discourse units. This problem was initially visualised by Āchārya R. M. K. Sinha and solved to a great extent with ANGLAB-HARTI approach in Indian context [5] for Hinglish that accepts input in the form of a variant discourse texts. The problem of discourse language attracted more attention while working with the news texts, where the headlines contain purely discourse sentences [8]. A number of strong constraints have been proposed that restrict the sequencing and attaching of segments at various descriptive levels, as well as the interpretation of their interrelations. The discourse texts are to be analyzed in terms of psychological, social and cultural practices of the group unto the society level. The utterance level of the text need to be analyzed that specifies the importance and significance of the event. To analyze these significant aspects of the discourse text a special kind of annotated corpora is essential apart from logic. However, discourse texts can be analyzed at a wide range of domains rather than in general form of language utterances.

During the literature survey to work on this present area, it is noted that the theoretical research work towards Critical Discourse Analysis (CDA) is in its infancy. So far, to the best of my knowledge there is no Machine Translation system working towards CDA for the language English itself. Thus an attempt was made to explore the possibility of implementing CDA for the language English at this stage. The reason behind such idea is the availability of adequate lexical resources and knowledge in English.

This paper is intended to present a theoretical framework, aimed towards investigating the possibility of applying Indian Logic preached by Dignāga, for handling discourse sentences during MT. Section I-A presents the theory of Dignāga,

Kommaluri Vijayanand is with the Department of Computer Science, Ramanujan School of Mathematics and Computer Science, Pondicherry University, Puducherry-605014, India (e-mail: kvixs@yahoo.co.in)

emphasized on various forms of valid knowledge, followed by the study on CDA that supports solution to deal with the discourse sentences that appear in texts during MT in section II. Section III presents the architecture for applying Logic to the machine. Section IV exemplifies various lexical resources used and their implementation details, followed by the conclusions in section V.

### A. Knowledge and Analysis

The Nyāya praveśa is an excellent work on Logic by Dignāga. The core theme of this volume [7] is stated briefly as:

Demonstration and refutation together with fallacies are useful in arguing with others; and Perception and Inference together with their fallacies are useful for self understanding; seeing these I compile this Śāstra.

### B. Sources of Knowledge

Perception and Inference are the two kinds of knowledge for one's own self. Perception is the knowledge derived through the senses whereas Inference is the knowledge of objects derived through a mark. Perception of a thing consists of the knowledge of its individual characteristics alone. Knowledge derived through inference is general and can be well expressed by name, genus, etc., whereas, that derived through Perception is particular and is capable of being properly communicated to others by name, genus, etc.

Inference for one's own self is defined as the knowledge of a thing derived through its mark or sign (middle term) of three characters, viz. Effect, Identity, and Non-perception. An inference for the sake of others takes place when a person demonstrates to others the conclusion drawn by him through inference for one's self. The affirmative reason signifies that the thing signified by its invariably accompanied by the thing signified by the predicate. Another source of knowledge is Apoha, an entity being the negation of its opposite.

### C. Reasoning

The literature against Logic is found in Vedas (composed between 1500 B.C. and 600 B.C.) which are regarded to be the oldest records not only of India but of the whole Aryan world. The term "Nyāya" in the sense of Logic does not appear to have been used in literature before the first century A.D. Pāṇini (about 350 B.C.) did not know the word "Nyāya" in the sense of Logic [1], and even Patañjali (about 150 B.C.) does not seem to have been conversant with the word.

1) Nyāya śāstra: Nyāya śāstra is the science of true reasoning. According to [6], Nyāya is defined as an examination of objects by evidences. He takes evidences to signify a syllogism which consists of a 'proposition' based on verbal testimony, a 'reason' based on inference, an 'example' based on perception, an 'application' based on comparison, and a 'conclusion' based on all the previous four. Vātsyāyana (about 400 A.D.) uses the expression "parama-nyāya" for the conclusion (nigamana) which combines in itself all the five parts of a syllogism.

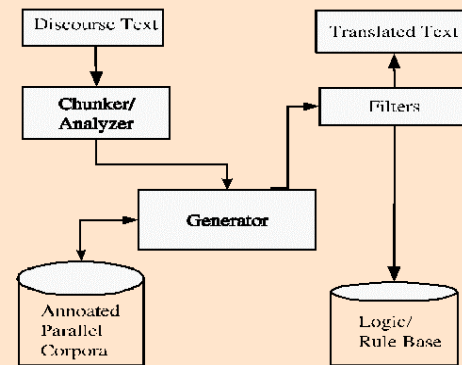


Fig. 1. Architectural Design for Discourse Translator

Dignāga (about 500 A.D.) explicitly mentions the five parts or members of a syllogism as Nyāyāvayava. In the Sabhāparva of the Mahābhārata we find Nārada was an expert in Nyāya śāstra. Nārada is said to be clever as he could ascertain the validity and invalidity of a speech of five parts.

2) Nyāya praveśa: Reasoning, according to the Nyāya praveśa, is carried out by means of a major term, a middle term, and two examples. The minor term also called the subject, the major term also called the predicate, and the middle term called the reason, combined together with two examples form the reasoning. A minor term and a major term combined together form a proposition, e.g., The hill (minor term) is ery (major term). The characteristics of the middle term are:

- The whole of the minor term must be connected with the middle term.
- All things denoted by the middle term must be homogeneous with the things denoted by the major term.
- None of the things are heterogeneous from the major term must be a thing denoted by the middle term.

Let us assume subject to be  $S$ , reason to be  $R$ , and predicate to be  $P$ . Now, we can generalize the characteristics mentioned above as shown in example 1. The negative aspect of the middle term only confirms the truth conveyed by one of the positive aspects, viz. All  $R$  is  $P$ . Hence, we can omit the negative aspects and exhibit the positive aspects as depicted in example 2. However,  $R$  and  $P$  may be taken in whole extent or partially. Hence, the positive aspects mentioned above may be fully exhibited as given in example 3.

- All  $S$  is  $R$ .  
All  $R$  is  $P$ .  
No  $R$  is non- $P$ .
- All  $S$  is  $R$ .  
All  $R$  is  $P$ .
- All  $S$  is all  $R$ .  
All  $S$  is some  $P$ .
- All  $R$  is all  $P$ .  
All  $R$  is some  $P$ .

Combining the aspects 3 and 4 together, we find that a syllogism may be of any one of the forms shown in

examples 5, 6, 7 and 8. Therefore, we can conclude that All is all  $P$ , and All is some  $P$ . Another interesting aspect is the relative extension of the middle term and major term. They show that the middle term is universally, invariably, or inseparably connected with the major term. This is called Vyāpti in Sanskrit.

(5) All is all  $R$  (Conclusion):

Because All is all  $R$ ,

All  $R$  is all  $P$ .

(6) All is some  $P$  (Conclusion):

Because All is all  $R$ ,

All  $R$  is some  $P$ .

(7) All is some  $P$  (Conclusion):

Because All is some  $R$ ,

All  $R$  is all  $P$ .

(8) All is some  $P$  (Conclusion):

Because All is some  $R$ ,

All  $R$  is some  $P$ .

(9) The hill is fier,

Because it has smoke,

All that has smoke is fiery as a kitchen.

(10) Whatever is not fier has no smoke as a lake.

An example can be converted into a universal proposition i.e., Vyāpti which stand to each other in the causal relation. The example 9, is homogeneous. The heterogeneous example may be laid down as shown in example 10.

## II. CRITICAL DISCOURSE ANALYSIS

Today, language and meaning are in some way social constructs. Emphasis on both the structure and the social context of media texts can provide a solution. CDA is leading to the development of a different approach to understanding media messages. The basics of a text consist of syntax and lexicon; its grammar, morphology, phonology, and semantics. However, the understanding of grammar and lexicon does not constitute the understanding of text. The comprehension of meaning lies not in the text itself, but in the complex interaction between the author's/receptor's intent and his/her performative ability to encode/decode that intent. CDA provides an opportunity



Fig. 2. Billboard at wash basin.

to adopt a social perspective in the cross-cultural study of media texts. Texts are selected and organized syntactic forms whose "content-structure" reflect the ideological organization of a particular area of social life. The language of media, as a particular style of discourse, is a complex blend of national, social, economic, and linguistic traditions which work in tandem with audience expectations. Discourse analysis employs the term in two broad categories of use [4]:

- 1) Discourse as an abstract noun denoting language in use as a social practice with particular emphasis on larger units such as paragraphs, utterances, whole texts or genres.
- 2) Discourse as a countable noun denoting a 'practice not just of representing the world, but of signifying the world, constituting and constructing the world in meaning'.

### A. Interpretation

Language cannot be separated from interpretation. Analysis need to be sensitive to their own interpretative tendencies and social reasons for them. Textual interpretation is psychological that is derived by the intelligence of a person. Such intelligence is applied to decode the text with the background information in understanding the text. Not only do different types of text require different ways of reading, but the same can also be read in different ways to generate different meanings. The relationship of the lexicon to the social context of the utterance can be thought of as exemplifying the way in which codified sign systems in general (verbal, visual, behavioural) are rendered meaningful only in relationship to the social structures which constitute them [2].

The texts that appear as discourse deal with the social structure of the society and need insight into the social and cultural aspects of the society. Recently, I had come across a billboard affixed by the eco-lovers of the society reproduced in the figure 2 and 3. The text appeared in the billboard contains text and images. As a Computational Linguist, I find only one valid sentence, i.e., "Water is Life". The next phrase

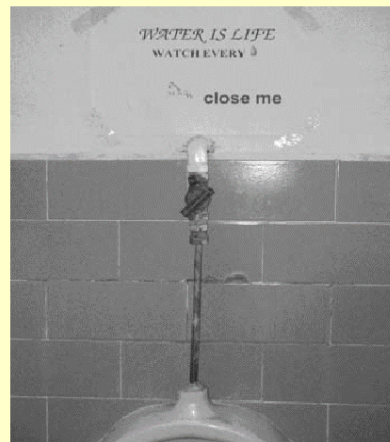


Fig. 3. Billboard at toilet.



contains “Watch every Drop”, where Drop is an image. The link between this sentence and the previous one is Water. The reader receives the knowledge from the previous sentence to understand the image of a “Drop of water”. Rest of the billboard requires some ‘common sense’ to understand. In addition to understand the message delivered, the location of the appearance of the billboard is very important and considerable. Though, the message appeared to be precise, it delivers a message of paragraph to the audience. When we compare the figure 2 and 3, the meaning one can derive can be same or different. If the reader do not possess any additional knowledge the meaning shall be the ‘same’.

### III. THE ARCHITECTURE

Based on the problems posed and analysis emphasized on understanding towards translating the discourse texts, we had designed an architecture. The core components of the proposed architecture are the Discourse analyzer, the Generator, and the Filters. The input to the system are discourse texts which commonly appear in news headlines. The chunker or the analyzer receives these input texts and tokenize them according to the pre-defined constraints that appear in the text like newline characters, size of fonts, images, etc. Based on these tokens, the Generator shall retrieve all possible translations, based on a specific approach viz., Rule-based, Statistical-based, Example-based or Hybrid-based approaches. The outputs from the Generator are received by the Filters, who shall apply the logic and context-based knowledge to discard the irrelevant translations. Thus, the sentences after filtering shall be the actual translations for the discourse texts.

#### A. The Analyzer

The chunker or sentence analyzer acts based upon the important aspects of the text received. The analysis of the text shall be composed with the Prosody, Cohesion, Discourse organization, Con-textualization signals, and Thematic organization. The system shall differentiate between ordinary pauses like comma, semicolon, hyphen, colon, etc. and the pregnant pause. The core theme is that, machine needs to identify the perceptiveness and authority of the writer. Verbal indicator of the prosodic features of stress is to be identified. This can be accomplished by identifying the words like long, long been suspected, barely, far, absolutely etc. This is an important task in identifying the writer’s argument towards his authority by a form of reiteration. Cohesive links are to be recognized which help in finding out the thread that tied the language and sense together. These cohesive links are conjunctions, pronouns, demonstratives, ellipses, adverbs and repeated words and phrases.

The rule-base contains a set of rules that will govern the behavior of the filters during the filtering process. Those strategies that can be adopted during implementation are as follow:

- 1) Choose the rules which best approximates the knowledge.
- 2) Design the rule to model the knowledge.
- 3) Learn the rules from examples generated successfully.

Applying these principles we implemented the machine translation system as explained in the section IV. The rule-base shall represent the domain specific knowledge.

#### B. Knowledge Representation

The logic that was discussed in section I-C can be represented using the Generalized Quantifiers (GQ). A GQ can be defined as: for a given set  $t$ , a GQ on  $t$  is a relation among subsets of relations on  $t$ . For example, When  $A$  universe  $t$  is fixed, we use  $Q$  as a variable over GQs,  $A, B, \dots$  as variables over sets, and write  $Q(A, B)$  to indicate that  $A, B$  belong to the extension of  $Q$ . That is, they are in the relation denoted by  $Q$ . However, few quantifier are similar to usual logic constants and behave independently towards the context.

**all** =  $\{A, B\}$

**some** =  $\{\text{at least } n\}$

**no** =  $\{\text{at most } n\}$

### IV. IMPLEMENTATION DETAILS

Towards implementing the system for CDA, what is essential is the domain specific knowledge. For example, the text that appeared in figure 3 would be perceived by a reader not to waste the water. Such understanding will be computed mentally with the knowledge that is acquired and stored in one’s memory. Therefore, unless the system/human possess the domain specific knowledge, it is impossible to understand the meaning. The domain specific knowledge that is involved in CDA shall be the knowledge about the culture, behavior, habits etc. The discourse texts do not contain the complete sentence, complete meaning, complete message. Now-a-days discourse texts are becoming more popular in advertising. An example can be the advertisements for liquor brands. The message is directed just to the alcohol consumers bypassing the minors. Hence, we need to train the system with a kind of knowledge to be aware of the general knowledge.

#### A. Lexical Resources and their Representation

Such training is essential to transform it to be intelligent, we need lexical resources where the system shall rely upon. Hence, we had attempted to teach the system with the knowledge-base created using a set of rules. These rules are collected from the reasoning exercises and trained the system using Prolog. The examples given in 11, 12, 13 and 14 are some of the forms of rules that are acting as knowledge base to the system, which can guide the system to conclude after its reasoning process. In all circumstances, the rule 11 shall treat that **All human are mortal** or **All men are human**. Like wise the experiment was successful to generate the conclusions for any given input with the help of these rules. The knowledge base can be improved to any extent according to the users requirement. However, at this stage the experiment was carried out to the extent of predicate logic with  $\forall$  construct.

(11) **mortal(X):- human(X)**  
**human(men).**

(12) **tasty(X):- sweet(X)**  
**sweet(rasagolla).**

- (13) **power**  $X$ ):- **rangers**  $X$   
**rangers**(spd).  
(14) **good**  $X$ ):- **faithful**  $X$   
**faithful**(dogs).

### B. Methodology

The system was implemented using the Rule-based methodology where the input is tokenized and transferred to match with the knowledge-base. For every given successful match it returns the conclusions. The system will form a predicate for every set of rules that are stored in the knowledge-base. The reasoning is carried out by applying predicate logic on the rules. Thus the system shall derive conclusions for a given pair of predicate sentences. The following example 15 narrates the reasoning for the construct. Interface to Prolog was provided using Visual C++ for creating work space. Some of the screen shots are given in the appendix.

- (15)  $\forall X, X \text{ s } M \text{ f } X \text{ s } a$ .

### V. CONCLUSION

The paper attempted to present several issues in discourse analysis along with Indian Logic that may be used towards knowledge acquisition. Towards handling various texts in Indian languages. Implementation of this system shall answer to several unsolved problems in the area of Machine Translation. While implementing the present system, we had used the rules in general instead of domain specific knowledge. Though we had listed several general application areas of CDA, still this problem shall be useful towards computational requirements like on-line MT during chatting and conferencing. As media is the powerful channel that deliver information to large

masses, operationally one yield maximum throughput from the MT systems when delivered with minimum number of limitations. It was a major difficult task to analyze and design the present MT system due to its multidisciplinary nature. We were successful in implementing the present system with minimum utilization of resources and producing the initial results. The future enhancements to the system are to extend the work to deal with the predicate logic for **some of**, **some of** and **for all**. During the investigation of the present system it is observed that implementing the predicate logic **not** is impossible. Elaborate study is necessary to apply the principles and implementation details discussed towards Indian Languages.

## APPENDIX A SCREEN SHOTS OF THE SYSTEM

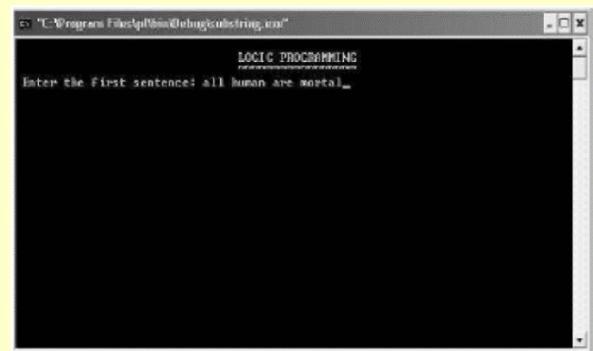


Fig. 4. Screen Shot to enter the input.

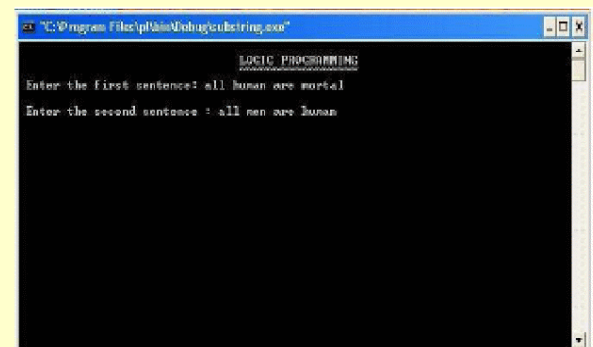


Fig. 5. Screen Shot after a pair of input.

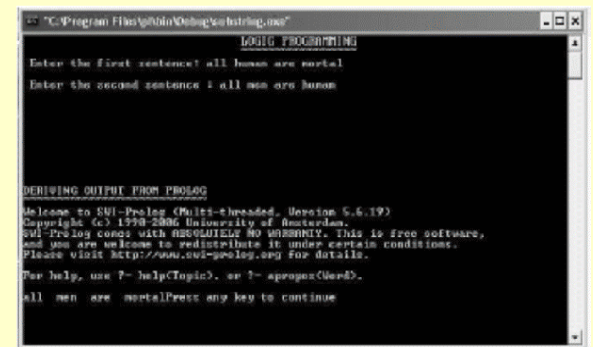


Fig. 6. Screen Shot to display the result.

## ACKNOWLEDGMENT

The author would like to dedicate this paper to his parents who were school teachers and filled their son with teaching interest. The author would like to thank Prof. S Kuppaswami, for his constant support extended and his colleagues of the Department.

## REFERENCES

- [1] Goldstucker, "Panini," p. 157.
- [2] T. Locke, Critical Discourse Analysis. London: Continuum International Publishing Group, 2004.
- [3] C. Moraga, F. Trillas, and S. Guadarrama, "Multiple-valued logic and artificial intelligence fundamentals of fuzzy control revisited," *Journal of Artificial Intelligence Review*, vol. 20, pp. 169–197, 2003.
- [4] A. Pennycook, *Critical Applied Linguistics: A critical introduction*. Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
- [5] R. M. K. Sinha and A. Thakur, "Machine translation of bi-lingual hindi-english (hingsh) text," in *Proceedings of the 10th Machine Translation Summit (MT Summit X)* Phuket, Thailand, September 13–15 2005.
- [6] Vatsyayana, *Nyāya-bhāṣya*, 1-1-1.
- [7] S. C. Vidyabhusana, *A History of Indian Logic*. New Delhi, India: Motilal Banarsidass, 1920.
- [8] K. Vijayanand and R. Subramanian, "Anuvadini: An automatic example-based machine translation system for bengali into assamese and oriya," in *Proceedings of the First National Symposium on Modeling and Shallow Parsing*, Indian Institute of Technology, Bombay, April 2–4 2006, pp. 267–281.
- [9] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Journal of Artificial Intelligence Review*, vol. 18, pp. 77–95, 2002.
- [10] D. Westerstaht, *Handbook of Philosophical Logic*. Dordrecht: Reidel Publishing Company, 1989.

*This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.*