AN INTELLIGENT SYSTEM FOR PICTURE BASED TAMIL SENTENCES GENERATION

Dr.T.Mala, Dr.T.V.Geetha
Department of Computer Science and Engineering
College of Engineering, Guindy, Anna University, Chennai

ABSTRACT

The existing picture dictionaries available electronically are static in nature. The user selects a picture and the contents related to the picture will be retrieved from the database and will be displayed on the screen. The drawbacks of the existing system are many. To mention a few, it is not intelligent, category of the picture is pre-defined, it is not dynamic, new sentences based on pictures cannot be generated and the semantic relationship between the pictures is not maintained. The proposed system aims at developing an intelligent system to generate the Tamil sentences automatically. Domain related pictures are provided. The user has to select more than one picture, based on the selection of the pictures; the semantic relationship between the pictures will be extracted from the semi-automatic domain ontology. Picture words and the semantically related words are sent to the sentence structure framework to obtain the syntactic representation of the Tamil sentence. Then the suffixes are added to the words to generate syntactically, semantically and morphologically correct sentences in Tamil.

1.INTRODUCTION

In the proposed picture based sentence generation system, computer programs are made to produce high-quality natural language text from computer internal representations of information. The system generates automatically a meaningful, grammatical and well formed sentence(s) about the set of pictures on which the user points out. The sentence(s) are generated in Tamil. The main part of the entire system is the domain ontology construction. The corpus is given as the input for the automatic ontology construction subsystem. Automatic domain ontology construction involves two modules. They are ontological word selection and semantic relationship identification between the ontological words. Using the terms extracted from the above modules and the sentence structure ontology the sentence structure is identified and finally suffixes are added to the words to generate syntactically, semantically and morphologically correct Tamil sentences. The entire paper is organized as follows. Section 2 discusses the literature survey in the areas of ontology construction and natural language generation, section 3 gives the overall system

architecture, section 4 talks on automatic ontology construction, section 5 gives details related to sentence generation and section 6 gives the conclusion and future works.

2. LITERATURE SURVEY

Ontology construction itself is a big challenge. An extension of hierarchical-valued concept context is adopted to elicit concepts from original component descriptions to construct ontological hierarchy for functional concepts [1]. Another method constructs ontology semi automatically by delimiting documents for learning in the domain of pharmacy involving four steps [2]. Ontology can be constructed automatically from a set of text documents. If documents are similar to each other in content they will be associated with the same concept in ontology. Semantic relationship for ontology construction can be identified from the keywords which are extracted by the clumping properties of content – bearing words [3].

Natural language generation systems are to be studied for automatically generating the sentences. The more "principled" approaches to NLG often make use of a division of the problem into stages such as content determination, sentence planning and surface realization [4][5]. Some of the issues in automatic sentence generation for picture dictionary creation are given below:

- Automatic construction of ontology without any manual intervention.
- What words and syntactic constructions will be used for describing the content?
- How is it all combined into a sentence that is syntactically and morphologically correct?

The proposed system is constructed by tackling the above issues and the architecture of the overall system constructed is explained in next section.

3. SYSTEM ARCHITECTURE

The figure 3.1 represents the entire architecture of an intelligent system for picture based Tamil sentences generation.

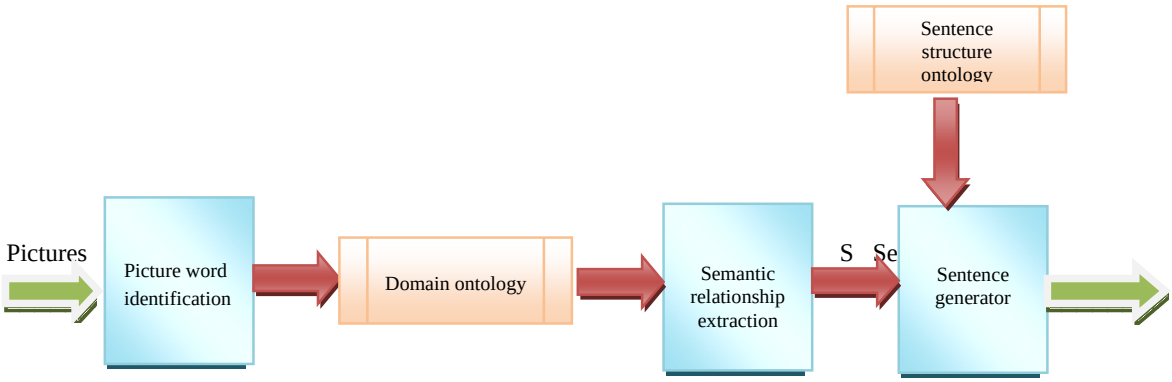

Figure 3.1 Overall system design

Users are intended to select pictures. Based on the selection of the pictures the corresponding picture words are identified. Since the preferred domain is animal domain, the input is restricted to the same. Picture words are searched in the animal domain ontology, which is already constructed semi-automatically as an offline process. The path is traversed from the bottom of the animal domain ontology to the top of it. The words thus extracted from the animal domain ontology are sent to the sentence structure framework to obtain the syntactic representation of the sentence. Finally suffixes are added to the words to generate morphologically correct sentences.

4. ONTOLOGY CONSTRUCTION

Tamil corpus forms the input to the system. Each and every Tamil text document is word segmented and morphologically analyzed to find out the parts of speech, by a tool called Atcharam (Tamil morphological analyzer). Then, the nouns are extracted and the noun list is confined by TF-IDF (Term frequency-Inverse document frequency) technique. Similarly semantically related words are extracted by the probabilistic framework and thus content bearing words are identified, from which the verbs are extracted to act as the semantically related terms. Using ontological node term and semantically related term, a domain ontology is constructed semi-automatically for Tamil text documents.

Semantic Relationship Extraction

Content bearing terms are identified using probabilistic framework from the text document corpus. Knowing the tendency of important terms to cluster serially, could be useful for extracting the semantic relationship of noun terms. Three measures are used to calculate the content bearing strength of a term and are given as term condensation over

textual units, term distribution over textual units and linear clustering. The tamil text documents are given as the input, to tag each and every ontological node term and semantically related term with their corresponding start and end tag. From the tagged tamil text documents, the connection between two ontological node with its semantically related term is identified.

5. AUTOMATIC SENTENCE GENERATION

To automatically generate the sentence the sentence structure frame work is to be defined first and then suffixes are added to the terms to generate syntactically and semantically correct sentences. Therefore the next stage is to construct the sentence structure frame work. There is no standard sentence structure for tamil. The following grammar rules were framed and based on the rule sentence structures are obtained [6].

1. NC --- > adj N / N / ADJC N / NNC
2. VC --- > adv V/ adv rpl / ADVC V / vpl / V
3. NNC --- > S con
4. ADJC --- > NC VC
5. ADVC --- > (NC)* vpl
6. S --- > (NC)* (VC)

Addition of suffixes was done using if – then rules. The next section gives the performance evaluation of the developed system and conclusion.

6. EVALUATION AND CONCLUSION

Extraction of ontological terms and semantic relationships are evaluated using that of experts and both gave an accuracy of about 80%. The accuracy of the generation of Tamil sentences is calculated by simple string accuracy formula which is defined in the below equation.

$$\textbf{Simple String Accuracy} = (1 – (I+D+S)/R)$$

where, I is the number of insertions, D is the number of deletions, S is the number of substitutions and R is the total number of tokens in the target string. It is inferred that as the total number of tokens gets increased, the accuracy gradually decreases and then gets increased. It proves that the system is efficient, since it can handle any number of tokens without affecting the accuracy. Thus the system generates syntactically, semantically and morphologically correct sentences. It also proves that the system is efficient, by comparing

the results of the system with the results of an expert and also by calculating string accuracy. In future, the project work can be extended to access the picture directly without using any picture word identification. This picture input can be chosen from the pictures provided in the user interface or the input can be given directly by the user. Also, a speech engine can be incorporated to convert the text to speech. Additionally, the user interface can be modified in such a way that it is also applicable for the physically challenged users.

REFERENCES:

1. Xin Peng, Wenyun Zhao, "*An Incremental and FCA-based Ontology Construction Method for Semantics-based Component Retrieval*", IEEE Seventh International Conference on Quality Software (QSIC), 2007.

2. Mu-hee song, Soo yeon lim, Ki-jun son and sang joo lee , " *Domain ontology construction based on semantic relation information of terminology*", IEEE industrial electronics society, November 2-6 2004.

3. Bookstein. A, Klein S.T, and Raita.T, "*Clumping properties of content bearing words*", IEEE proceedings of the sixth international conference on machine learning and cybernetics Hong kong, 19-22 August 2007

4. Auxilio Medina, Alberto Chavez-Aragon, "*Construction, Implementation and Maintenance of Ontologies of Records*", Proceedings of the Fourth Latin American Web Congress (LA-WEB'06), 2006.

5. Ehud Reiter, "*Building Natural Language Generation Systems*", 1996

6. Saravanan, K., Ranjani parthasarathi, Geetha.T.V., "*Syantactic Parser for tamil*", Tamil internet, 2003.