CONFERENCE PAPERS
Abstract

Summarization of large text documents becomes an essential task in many Natural Language processing (NLP) applications. Certain NLP applications deal with domain specific text documents and demand for a domain specific summary. When the essential facts are extracted specific to the domain, the summary proves to be more efficient. The proposed system builds a bilingual summary for an Information Retrieval (IR) system named CoRee, which tackles Tamil Language and English Language text documents [1]. As the input documents are tourism domain specific documents, the summary is extracted based on specially designed seven tourism specific templates 7 both for Tamil and English. The templates are filled in with the required information extracted from the UNL representation and a bilingual summary is generated for each text document irrespective of the language of input text document. The efficiency of the summary has been tested manually and it has achieved 90% efficiency. This efficiency depends on factors other than summary generation such as conversion accuracy and dictionary entry coverage. The proposed system can be extended for many languages in future.

1. Introduction

Automatic summary generation has been a research problem for over 40 years [2]. Summarizing the texts helps in avoiding information overload and also saves time. Multi lingual Natural Language applications have emerged in great number in recent years. This makes the need for a multi lingual summary generation a quintessential task. Alkesh patel et al have come up with a multi lingual summary generation by using structural and statistical factors [2]. David Kirk Evans has generated multi lingual summary using text similarities existing in the sentences [3]. Dragomir Radev et al have developed a multi lingual summary generation tool named MEAD using centroid and query based methods. They have also used many learning techniques such as decision trees, Support Vector Machines (SVM) and Maximum Entropy [4].

All the above works on multi lingual summarization have not used a interlingua document representation. We propose that a multi lingual summary can be generated with much more ease by using a interlingua document representation language called, “Universal Networking Language” (UNL) [5]. UNL converts every term present in a natural language text document into a language independent concept, thereby making the applications built using it a language independent one. The proposed work extracts a domain specific summary, as the UNL documents used are tourism domain specific. Tourism specific templates are framed and the sentences fitting the templates are chosen and formed as a summary.
The rest of the paper is organized as follows. Section 2 gives a brief introduction about UNL. Section 3 describes the proposed summarization technique. Section 4 discusses the evaluation of the proposed work. Section 5 reveals the enhancements needed to the proposed work and Section 6 gives the conclusion of the paper.

2. Universal Networking Language

UNL is an intermediate language that processes knowledge across language barriers. UNL captures the semantics of the natural language text by converting the terms present in the document to concepts. These concepts are connected to the other concept through UNL relations. There are 46 UNL relations like plf(Place From), plt(Place To), tmf(Time from), tmt(Time to) etc [1]. This process of converting a natural language text to UNL document is known as Enconversion and the reverse process is known as Deconversion. The UNL document is normally represented as a graph where the nodes are concepts and edges are UNL relations. An example UNL graph is shown for the example 1.

Example 1: John was playing in the garden.

![UNL graph for Example 1](image)

The nodes of graph namely, “John(iof>person)”, “Play(icl>action)” and garden(icl>place) represent the terms John, playing and garden present in the example 1. The semantic constraints in the concepts, “iof>person”, “icl>action” and “icl>place” denotes the context in which the concepts occur. The edges namely, “agt” and “plc” indicates that, the concepts involved are agents and place. From the above discussion, it is shown that the UNL inherits many semantic information from the natural language text and portrays in a language independent fashion.

The proposed work uses Tamil language text documents and English language documents enconverted to UNL for summary extraction which is described in the next section.

3. Template based Information Extraction

As discussed earlier, the summary is generated using the tourism specific templates. Figure 2 shows the over view of the proposed summary generation framework. The Framework consists of both language dependent and independent parts. The functionalities involving UNL are language independent and the inputs supplied to the framework to generate bilingual summary are the language dependent parts. The bilingual summary generation is explained in the coming sections.
The seven templates describe about the tourism specific information of a place such as, god, food, flora and fauna, boarding facility, transport facility, place and distance. The correct information for these templates are extracted as discussed below. The usage of semantics helps greatly in eliminating the ambiguities that may arise while picking up a concept to fill the slot. For instance, the word, “bat” may denote a cricket bat or the mammal bat.

Figure 2: Overview of the Summary Generation Framework

This type of ambiguity is resolved by the semantic constraint, as the cricket bat will get the semantic constraint, “obj<thing (object thing)”, whereas the mammal gets the semantic constraint, “icl>mammal”. Table 1 displays few semantics used for the respective templates.

The extracted tourism specific concepts are converted to the target language terms for building a summary using the sentence patterns which is explained in the next section.

4 Multi Lingual Summary generation

The information (concepts) extracted from the UNL graph using the templates are converted to the target language term using the respective UNL dictionary. For instance, to generate the English summary, the concepts comprising the semantic constraints are converted to English terms using the English UNL dictionary which consists of mapping between English terms and UNL concepts. These terms which when filled into the appropriate English sentence patterns, gives a English summary. The same procedure is done for building a Tamil summary. For each UNL graph irrespective of its source language, a summary in Tamil and English are generated.
<table>
<thead>
<tr>
<th>Template</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>God</td>
<td>iof&gt;god, iof&gt;goddess, icl&gt;god</td>
</tr>
<tr>
<td>Food</td>
<td>icl&gt;food, icl&gt;fruit</td>
</tr>
<tr>
<td>Flaura and Fauna</td>
<td>icl&gt;animal, icl&gt;reptile, icl&gt;mammal, icl&gt;plant</td>
</tr>
<tr>
<td>Boarding facility</td>
<td>icl&gt;facility</td>
</tr>
<tr>
<td>Transport facility</td>
<td>icl&gt;transport</td>
</tr>
<tr>
<td>Place</td>
<td>icl&gt;place, iof&gt;place, iof&gt;city, iof&gt;country</td>
</tr>
<tr>
<td>Distance</td>
<td>icl&gt;unit, icl&gt;number</td>
</tr>
</tbody>
</table>

Table 1: Semantics used for each templates

The terms obtained from the UNL dictionary will be a root word. For instance, the term, “eating” will be entered as eat (icl>action) in the UNL dictionary. So the terms obtained from the UNL dictionary needs to be generated to its original form using Morphological generator. The summary generation requires only tourism specific concepts, so the generation is almost not required. But we have used a morphological generator for Tamil, as the place information and distance information in Tamil with the case suffixes இ᾿ (il), இᾢᾞᾸᾐ (ilirunthu), உᾰᾁ (ukku) etc needs to be generated. For the example UNL graph shown in figure 3, the generated transport template in Tamil which is part of the summary is given in example 2.

Example 2: சென்னை (iof>city) போக்கு (icl>action)

The concept chennai (iof>city) in the above graph, is generated as "சென்னை" by adding the case suffix "இ᾿" and the concept bus (icl>vehicle) is generated as "பொக்கு" by adding the case suffix, "இ᾿".
5. Performance Evaluation

The proposed work has been tested with 33,000 Tamil and English text documents enconverted to UNL graphs. The performance of the methodology proposed has been evaluated using human judgement. The accuracy of the summary generated has achieved 90%. Apart from the summary generation factors such as tourism specific concept extraction, the accuracy also depends on the quality of enconversion and dictionary entry. By improving these factors, the accuracy can further be improved.

6. Conclusion and Future work

The proposed work generates a tourism specific bilingual summary using the intermediate document representation, UNL and tourism specific templates. The bilingual summary is generated in a simple and efficient manner compared to the earlier work done for multi-lingual summary generation. The only over head involved is developing a enconverter framework.

As future enhancements, sentence patterns can be replaced by selecting the sentences having high sentence score based on its sentence position and the frequency of concepts. Query specific summary can also be generated on line, as the summary discussed here is a tourism specific generated off line using the templates. The evaluation of the generated summary can also be done by comparing it with the human generated summary. By doing this, many factors to make the machine generated summary compatible with human generated summary may evolve.

Reference

- Radev, Allison, Blair-Goldensohn et al (2004), MEAD - a platform for multidocument multilingual text summarization
- Prof. R. Nedunchelian, “Centroid Based Summarization of Multiple Documents Implemented using Timestamps” First International Conference on Emerging Trends in Engineering and Technology, IEEE 2008