5.1 About Telugu

Telugu is one of the ancient Indian languages belonging to the Dravidian language family with its own unique script and is the official language of the state of Andhra Pradesh in south India. More than half of the Telugu vocabulary is from Sanskrit. It is the second largest spoken language in India after Hindi with a population of about 80 million people, close to the number of Korean or Vietnamese native speakers. Telugu has rich agglutinative characteristics. For example, a study done on the number of unique words found in the same size of corpus for various Indian languages showed that Telugu had the most number of unique words in the corpus than any other Indian language. This phenomenon is observed due to the agglutinative properties of the language. This feature of the language motivated us to choose Telugu as the topic language over other Indian languages [87].

5.2 Background

The first cross language efforts from Telugu to French and Telugu to English began in the year 1824 by an Englishman, Charles Philip Brown. During his 60 years of efforts in Telugu, Charles wrote many books in Telugu including Telugu to English and English to Telugu cross language dictionaries in the year 1854 which are still popular
among Telugus. These dictionaries were later digitized as part of Telugu - English resource collection for machine translation applications. These dictionaries were however written for human consumption and hence contained descriptive meanings for words. Therefore the first challenge is to convert them to be usable for cross lingual applications. These dictionaries also contain a list of parallel sentences between English and Telugu, which are usually provided as examples of word usage in cross lingual dictionaries, which can be useful for cross language information retrieval. However these sentences cannot be treated as a corpus in the statistical sense since they are not a statistical representative sample from general text such as news articles or any other text documents [87].

5.3 How the Telugu Shaping Engine Works

The Telugu shaping engine works on a Telugu Unicode character string in the following stages:

1. Identifies syllables
2. Reorders characters within the syllables
3. Shapes glyphs using OTL services
4. Positions glyphs using OTL services

First the shaping Engine identifies the boundaries of syllables and isolates certain parts. It does this to be able to place syllable elements in required positions and to determine if combining elements have to be
preferentially replaced by graphically distinct forms, such as the syllable kssa (క్ష). Then the base consonant, which is displayed in full form, is found, and all other elements are classified by their position relative to this base. These other elements could be pre-base (non-existent in Telugu), below-base (quite common), above-base (quite common) and post-base (quite common). The shaping engine splits the matras that occur on more than one side of the base consonant into their constituents. The phonetic sequence of Unicode characters initially passed on to Uniscribe may not be in the order that it can work with. So, Uniscribe reorders this sequence and maintains a buffer of such appropriately reordered character codes, delineated as "clusters". It then gets the appropriate glyph substitutions for these clusters where necessary, using OTL services. Since Uniscribe uses reordered character sequences to render the font on the screen, it is essential that the font designer take this behavior into account when designing the font. This re-ordering is done according to the following rules:

1. The base consonant is found by the following algorithm: starting from the end of the syllable move backwards until a consonant is found that does not have a below-base or post-base form (post-base forms have to occur after below-base forms), or arrive at the first consonant. The consonant stopped at will be the base consonant.

2. If the base consonant is not the last one, then Uniscribe moves the halanth from the base consonant to the last one.

3. Uniscribe splits two part matras such as AIMatra into above base and below base components.
In Telugu there is only one such case, i.e., AI Matra (₹). This is split into ₹-i.e., into an above-base E Matra (code 0C46) and a below-base AI Length Mark (code 0C56). For Telugu this classification is as follows:

**Below-base Markers:** All consonant markers (i.e., Ottulu), AI length mark, Anudatta. And the matras corresponding to Vocalic L and Vocalic LL. Note that the latter three markers are not yet defined in the Telugu Unicode range, which they must be. They are mentioned here for completeness.

**Above-base Markers:** The matras corresponding to vowels ə, ə, œ, ø, ɔ, ɔ, ʌ, ʌ, Udatta, Double Udatta, and Nukta. Note that Udatta, Double Udatta and Nukta are not yet defined in the Telugu Unicode range, which they must be. They are mentioned here for completeness.

**Post-base Markers:** The matras corresponding to vowels ə, ə, œ, ø, ɔ, ɔ, i.e., Kommu, Kommu-deerghamu, Rutvamu, Rutvam-deerghamu) are post-base matras. Although the Anuswara (Sunna), Ardhanuswara (Arasunna), Visarga and Avgaraha occur as post-base characters, they do not affect the shape of the base consonant or ligature, so they are classified as Vowel Modifiers (VMs), not as post-base matras. Note that Avagraha is not yet defined in the Telugu Unicode range, which it must be. It is mentioned here for completeness. There are no pre-base forms in Telugu. Uniscribe then groups elements of the syllable (consonants and matras) according to this classification. Thus pre-base elements precede the base
consonants, and the above-base, belowbase and post-base components will follow the base glyph. Halanths are moved with the consonants they affect. These principles are illustrated by the following example: Consider the Sanskrit word

\[ \text{लक्ष्मी} \]

This word has 2 syllables \( \text{ला} \), \( \text{क्ष्मी} \).

The first syllable is very simple consisting of a single base consonant. So it doesn’t need any rearrangements, substitutions or relative positioning [88].

### 5.4 Experimental Results

The experiments are done by taking the corpus, which was developed by the articles from Eenaadu newspaper over 100 days and converted to the font-encoded pages into ISCII standard encoding. The corpus includes 1970 articles totaling to 4.5 Million words. Of the 1970 documents in the Telugu News Articles corpus, 300 documents in 4
major categories (P-Politics, S-Sports, E-Editorial, and C-Cinema) have been used in the current set of experiments. The distribution of the documents across these four categories for each language is tabulated below.

Table 5.1 Distribution of the documents in Eenaadu News Articles

<table>
<thead>
<tr>
<th>Category Wise Break up</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Politics</td>
</tr>
<tr>
<td>300</td>
<td>100</td>
</tr>
</tbody>
</table>

It may be observed that the distribution of documents in various categories is not uniform - Cinema for example, includes fewer documents.

Below figure shows the comparison between the Human and Machine generated summaries for the various articles distribution shown below.

<table>
<thead>
<tr>
<th>News Category</th>
<th>Number of Articles</th>
<th>Machine</th>
<th>Number of Articles</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>01</td>
<td>92.18</td>
<td>100</td>
<td>89.12</td>
</tr>
<tr>
<td>Sports</td>
<td>01</td>
<td>91.12</td>
<td>80</td>
<td>90.12</td>
</tr>
<tr>
<td></td>
<td>01</td>
<td>90.34</td>
<td>70</td>
<td>88.15</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td><strong>Editorial</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cinema</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>91.22</strong></td>
<td></td>
<td><strong>89.16</strong></td>
<td></td>
</tr>
</tbody>
</table>

Fig 5.1: Comparison between Human and Machine generated summaries.