Morpheme Extraction Task at FIRE 2014

Nilotpala Gandhi\textsuperscript{1}, Kanika Mehta\textsuperscript{2}, Prashasti Kapadia\textsuperscript{2}, Adda Roshni\textsuperscript{2}, Vaibhavi Sonavane\textsuperscript{2} and Prasenjit Majumder\textsuperscript{2}

\textsuperscript{1}Gujarat University, Ahmedabad, Gujarat  
\textsuperscript{2}DA-IICT, Gandhinagar, Gujarat

Abstract. The Morpheme Extraction task (MET) was organized for the third time this year after introducing it in FIRE 2012. Participating systems were required to provide morphemes of given term lists. The track was offered in three languages viz., Bengali, Gujarati and Tamil. The evaluation exercise comprised of a linguistic evaluation for the submitted systems. This overview paper describes the goals, data, tasks, participants, evaluation process, and obtained results.

1 Introduction

Morphology is the study of words and their grammatical structure. Simply put, it is actually the study of `morphemes', the basic meaningful units of a language which combine to form more complex words. When morphemes exist freely on their own, they are known as the root words. Those morphemes which do not have an independent existence are known as bound morphemes or axes.[1] For example, the word `sleeping' is made of two morphemes: sleep and -ing. Here sleep is a free morpheme while -ing is a bound morpheme. The goal of morphological analysis could be either to understand the structure of a language and use this understanding in various interesting tasks in machine translation, natural language processing etc. or to simply improve the retrieval performance in that particular language. Extracting morphemes from a language many times involves an important process called `Stemming' or affix removal. Three different kinds of stemmers are found in literature. The first kind, known as supervised stemmers, depend on known grammatical rules of a language whereas, the second kind, i.e., unsupervised stemmers are statistical, algorithmic and do not need any language-specific information. The third kind lie between the first two and are known as semi-supervised stemmers.[2]

Many Indian languages are morphologically rich because of the presence of huge amount of different word forms. The vast amount of various in inflected forms in which the words appear, poses an important challenge for information retrieval experiments in Indian languages. Hence, morpheme analysis becomes an important step for Information retrieval in Indian languages.
In MET 2014, both language dependent and language independent systems are evaluated for Indian languages. Linguistic evaluation was carried out for the submitted systems. These, along with the results obtained, will be discussed in detail in the following pages. The task was closely modeled based on the Morphochallenge 2010 conducted by Department of Computer Science, Aalto University, Finland. [3] (http://research.ics.aalto./events/morphochallenge2010/)

2 Data

The FIRE ad hoc corpora were used for all languages. The word list for each language has been constructed by collecting all the terms that occur in the corpus after filtering it of English terms, numerical digits, punctuation marks etc. The term lists were made available on the FIRE website.

Some of the statistics about the data used are shown in Table 1.

<table>
<thead>
<tr>
<th>Language</th>
<th>No. of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td>1315518</td>
</tr>
<tr>
<td>Gujarati</td>
<td>1932607</td>
</tr>
<tr>
<td>Tamil</td>
<td>666683</td>
</tr>
</tbody>
</table>

Table 2 shows the number of words of gold standard data used in Bengali and Tamil for linguistic evaluation.

<table>
<thead>
<tr>
<th>Language</th>
<th>Training Data</th>
<th>Test Data</th>
<th>Total No. of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td>600</td>
<td>1074</td>
<td>1674</td>
</tr>
<tr>
<td>Tamil</td>
<td>1000</td>
<td>1938</td>
<td>2938</td>
</tr>
</tbody>
</table>

3 Task

Linguistic Evaluation

A sample of the proposed morpheme analyses of the systems were compared against a sample of the gold standard data (which contains manual morpheme analyses). This experiment was repeated over several samples and the average was treated as the final score. This task was performed for Bengali and Tamil languages, as these are the only
languages for which some gold standard data is available.

While comparing the two analyses, it cannot be expected that the algorithm of the system comes with the morphemes that exactly correspond to the ones in the gold standard data. So word pairs with same morphemes are compared and scores are calculated on the number of matched thus obtained.

For obtaining Precision, a set of 1000 words are sampled from the result files generated by the morph analysis systems. For each of the sampled words, another word having same morpheme is chosen from the result file and these pairs are compared to the gold standard data. A point is giving for every word pair that has common morpheme in the gold standard as well. The number of points for each word is normalized to 1. Now, precision is calculated as the ratio of total number of points obtained to the total number of sampled words.

\[
Precision = \frac{\text{Number of points}}{\text{Total Number of sampled words from the result file}}
\]

Similarly for calculating Recall, a set of 1000 words are sampled, but this time from the gold standard data. For each of these words another word having the same morpheme is chosen from the gold standard data randomly. The word pairs are then compared to the analyses in the results generated by the morphanalysis systems. A point is giving for each sampled word pair having common morpheme. Recall is calculated as follows:

\[
Recall = \frac{\text{Number of points}}{\text{Total number of words from the gold standard data}}
\]

This process is carried out several times and the average values of Precision and Recall are taken. The final score for the system in this evaluation is taken as the F-measure i.e., the harmonic mean of Precision and Recall,

\[
F - measure = \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}
\]

The code, used for Morphochallenge 2010, which implements all the above steps is used for this evaluation.[3]
4 Participants

Although 4 participants registered for the task, only two of them submitted their run. Table 3 shows the participant names and the languages supported by their system.

Table 3. LIST OF PARTICIPANTS

<table>
<thead>
<tr>
<th>Institute</th>
<th>Participants</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISI Kolkata</td>
<td>Abhishek Chakrabarty, Sharod Roy Choudhary, Dr. Utpal Garain</td>
<td>Bengali</td>
</tr>
<tr>
<td>Amrita Vishwa Vidhyapeetham</td>
<td>Dr. M. Anand Kumar, Dr. K. P. Soman</td>
<td>Tamil</td>
</tr>
</tbody>
</table>

5 Results

Table 4 shows the Precision, Recall and F-measure scores for ISI's stemmer and Amrita University's Tamil morphological analyser. The scores have been computed as described in the Task section.

(The evaluation of Amrita University’s Tamil morphological analyser was done on 500 words so far. So the following results are for 500 (out of 2938) words only)

Table 4. RESULTS

<table>
<thead>
<tr>
<th>System</th>
<th>Language</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISI Kolkata</td>
<td>Bengali</td>
<td>56.19%</td>
<td>65.08%</td>
<td>60.31%</td>
</tr>
<tr>
<td>Amrita Vishwa Vidhyapeetham</td>
<td>Tamil</td>
<td>10.86%</td>
<td>40.90%</td>
<td>17.16%</td>
</tr>
</tbody>
</table>

ISI's system performs pretty well on recall as well as on precision. The results from Amrita University aren't very satisfactory but the analysis on the complete test data is yet to be done.

6 Conclusion

Morpheme Extraction task successfully evaluates some of the latest systems for stemming and morphological analysis in Indian languages. Its main goal has been to encourage participants to experiment with different methods to improve their systems and obtain better scores. ISI's system has shown some improvement as compared to previous year's results. The results from Amrita University aren't very promising so far but better results are expected when tested on the larger test data. The improved systems will be useful in information retrieval, text understanding, machine learning and language modeling.
7 Acknowledgment

We are grateful to the team of FIRE organizers and the IR lab team at DA-IICT for their support and assistance. We especially wish to thank Parth Mehta and Ayan Bandyopadhay for their invaluable help in maintaining the MET webpage. The gold standard analyses data provided by IIT-Kharagpur and AUKBC is also gratefully acknowledged. Lastly, we also thank all the participants of this task for their enthusiasm in submitting their system.

References