Morpheme Extraction Task at FIRE 2013

Rashmi Sankepally[1], Komal Agarwal[2], and Prasenjit Majumder[2]

[1] University of Maryland, MD, USA
[2] DA-IICT, Gandhinagar, Gujarat

Abstract. The Morpheme Extraction task (MET) was organised for the second 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 six languages viz., Bengali, Gujarati, Hindi, Odia, Marathi, Tamil. The evaluation exercise comprised of retrieval and running of a given corpus with a standard set of queries with fixed relevant judgements and culminated by finding the Mean Average precision for the runs. This year, linguistic evaluation was also done for Tamil and Bengali morph analysis 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 affixes.[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 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 2013, both language dependent and language independent systems are evaluated for Indian languages. Two different kinds of evaluation methodologies were followed. 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. [5] (http://research.ics.aalto.fi/events/morphochallenge2010/)

2 Data

The FIRE adhoc corpora were used for all languages. The corresponding Queries and Qrels were used. Stop words are also used from FIRE data.

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 1. DATA STATISTICS

Language	Corpus	no. of terms	corpus size	no. of docs	Query no.
Bengali	FIRE 2011	1359816	3.4 GB	500122	126 - 175
Hindi	FIRE 2011	407559	2.0 GB	331559	126 - 175 (v2)
Gujarati	FIRE 2011	1932864	2.8 GB	313163	126 - 175
Marathi	FIRE 2010	861063	694 MB	99275	76 - 125
Odia	FIRE 2012	222535	208 MB	32872	176 - 225
Tamil	FIRE 2011	666683	1.02GB	194483	176 - 125

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

Table 2. Gold Standard Analysis Data

Language	Training Data	Test Data	Total No. of words
Bengali	600	1074	1674
Tamil	1001	1938	2939

3 Task

3.1 IR Evaluation

The morpheme analyses proposed by the systems submitted were used in Terrier indexing and retrieval experiments. The retrieved results were evaluated against the available relevance judgments. The Mean Average Precision thus obtained was treated as the final score for the system. This evaluation was made available for Bengali, Gujarati, Hindi, Marathi and Odia languages. Terrier-3.5 was used to perform all the runs and evaluation was done with Trec-Eval-9.0.[7]

3.2 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 morpheme 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.
$Precision = \frac{Numberofpoints}{Totalnumberofsampledwordsfromtheresultfile}$

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 morph

analysis systems. A point is giving for each sampled word pair having common morpheme. Recall is calculated as follows:
$Recall = \frac{Numberofpoints}{Totalnumberofsampledwordsfromthegoldstandarddata}$

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,
$F - measure = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$

The code, used for Morphochallenge 2010, which implements all the above steps is used for this evaluation.[5]

4 Participants

Unlike the participation it recieved last year when it was first introduced, MET has received luke warm participation this year. Although 5 participants have registered to participate, only 3 participants have actually submitted their runs. Table 3 shows the participant names and the language(s) supported by their system.

Table 3. List of Participants

S.No.	Institute	Participant Names	Language(s) supported
1	ISM-Dhanbad	Amit Jain, Nitish Gupta	language independent
2	AUKBC, Chennai	Sobha Lalitha Devi, Marimuthu K, Vijay Sundar Ram R, Bakiyavathi T, Amudha K	Tamil
3	MANIT, Bhopal	Anubha Jain, Sujoy Das	Hindi

5 Results

Table 4 5 : Show the Mean Average Precision(MAP) values obtained for different languages by the systems submitted. The MAP for Basline run (run with no stemming) is also given. Fourth column contains %improvement over the baseline score.

Clearly both the systems perform better than the baseline score for most languages except ISM's system for Hindi. ISM's language independent system performs exceptionally well with Bengali and Marathi.

Table 4. ISM Results

Language	Baseline MAP	MAP obtained	% improvement
Bengali	0.2740	0.3158	15.25%
Hindi	0.2821	0.2793	-0.99%
Gujarati	0.2677	0.2824	5.49%
Marathi	0.2320	0.2797	20.56%
Odia	0.1537	0.1583	2.99%

Table 5. MANIT Results

Language	Baseline MAP	MAP obtained	% improvement
Hindi	0.2821	0.2917	3.40%

Table 6 shows the Precision, Recall and F-measure scores for AUKBC's Tamil morpheme analyser and ISM's system for Tamil and Bengali languages. The scores have been computed as descibed in the Task section.

Table 6. Linguistic Evaluation Results

System	Language	Precision	Recall	F-measure
AUKBC	Tamil	84.29%	88.15%	86.17%
ISM	Tamil	80.22%	18.86%	30.54%
ISM	Bengali	60.64%	32.15 %	42.02%

Results from linguistic evaluation indicate that AUKBC's system shows significant accuracy in discovering morphemes and their affixes as evident from its high precision and recall scores. ISM's system performs well on precision but not as good with recall which implies that it could not accurately map sampled pairs with same morphemes from the gold standard data. It is to be noted that ISM's system has not been tested on affixes as it only produces root morphemes. The results in linguistic evaluation are constrained by the small sizes of the training and testing data.

6 Conclusion

Morpheme Extraction task could successfully evaluate 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. Most of the systems submitted show significant improvement over baseline scores, implying their usefulness in information retreival, text understanding, machine learning and language modeling.

7 Acknowledgements

We are grateful to the team of FIRE organizers and the IR lab team at DA-IICT for their support and assisstance. We specially 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 ackowledged. Lastly, we also thank all the participants of this task for their enthusiasm in submitting their systems.

References

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: An Introduction to Information Retrieval, Online edition (c) 2009 Cambridge UP, Draft of April 1, 2009.
2. Harald Hammarström, Radboud Universiteit and Max Planck Institute for Evolutionary Anthropology; Lars Borin, University of Gothenburg: Survey Article- Unsupervised Learning of Morphology,(c) 2011 Association for Computational Linguistics, 4 October, 2010.
3. FIRE Website: http://www.isical.ac.in/ fire/2012/index.html
4. MET Website: http://www.isical.ac.in/ fire/morpho/MET.html
5. Mikko Kurimo, Sami Virpioja and Ville T. Turunen: PROCEEDINGS OF THE MORPHO CHALLENGE, TKK-ICS-R37, TKK Reports in Information and Computer Science Espoo 2010, Aalto School of Science and Technology http://research.ics.aalto.fi/events/morphochallenge2010/
6. PRASENJIT MAJUMDER, MANDAR MITRA, SWAPAN K. PARUI, and GOBINDA KOLE Indian Statistical Institute PABITRA MITRA Indian Institute of Technology and KALYANKUMAR DATTA Jadavpur University: YASS: Yet Another Suffix Stripper, ACM Transactions on Information Systems, Vol. 25, No. 4, Article 18, Publication date: October 2007.
7. TREC(Text REtrieval Conference website): http://trec.nist.gov/