Workshop on Multi Word Expression
April 11th, 2008
AU-KBC Research Centre, MIT Campus, Chennai

***Institutes Participated*:** AU-CEG, CDAC Noida, CDAC Pune, IIIT Hyderabad, IIT Bombay, IIT Kharagpur, HCU, Jadavpur University, Tamil University ,and AU-KBC.

***Workshop Conductors*:** Anoop Kunchukuttan and Prof Om Damani from IIT Bombay.

Morning Session:

- Definitions, Motivation, and Characteristics of Multiwords were discussed.
- It became clear in the discussion that classifying multiwords based on the given definition and characteristics is a hard and error-prone task. Hence a set of operational criterions were developed, i.e., a set of rules in priority order were given for identifying MWEs.
- IIT Bombay promised to document and distribute these guidelines in a week's time. (Being released along with these minutes.)
- It was emphasized that while deciding whether a word group was MWE or not, one should always refer to the guideline and specify the rule using which something is marked MWE. Sometimes, we may intuitively feel that something is an MWE but none of our rules may apply. In that case either the rules or the intuition will get corrected.
- It was emphasized that we were only concerned with space or hyphen separated group of words. Hence we were ignoring *Samasik* words where more than one words get glued together to form a single word. For the current purpose, we assume that all such *Samasik* words will be listed in the lexicon.

***Afternoon Session*:**

- It was decided to focus on the Noun-Noun MWEs in the beginning since they are the most important ones for the CLIA purpose. Once we get sufficient experience with NN MWEs then we will move to other MWEs.

- An overview of approaches to Multiword Extraction, especially of the system developed at IIT, Bombay was discussed. The tool currently handles only the Noun-Noun MWEs. Given a POS Tagged corpus in the SSF format, the tool extracts all NN MWEs in a list, sorted by some scoring function. It also outputs MWEs sorted by their frequency, LLR (Log Likelihood Ratio), and PMI (Pointwise Mutual Information) scores.
  Given a 'gold-standard' list of MWEs for the same corpus, the tool also gives precision and recall scores.
- The original plan of running the tool on corpus for various languages could not be executed because the tool was not ready. Instead a demonstration on Tool usage using the Hindi corpus from IIT Bombay was given.
- IIT Bombay promised to distribute the tool in a week's time. (Being released along with these minutes.)
- There was further discussion on Evaluation Guidelines for the identification of multiword expressions.
- The team from Tamil University presented many MWE examples for Tamil. There was a discussion to see whether these terms were MWE in other languages too. It was found that except one example, all other terms were MWEs in other Indian languages too.

***Future Work*:**

- Multiwords to be manually extracted for entire 2,50,000 word Tourism corpora by 15 Jun 2008. In the first stage, MWEs to be extracted from 150,000 words corpus by May 15. MWEs should also be categorized according to criteria developed. This extracted list will be the 'Gold Standard', against which the tool can be evaluated.
- Another workshop on MWE will take place between June 15 and June 20 at AU-KBC.

Resources:

- MWE Tool
- MWE Extraction Guidelines
- MWE Storage Template
- Presentations

All these resources can be downloaded from:
http://www.cfilt.iitb.ac.in//~consortia/dokuwiki/doku.php?id=clia:multiword_expression_extraction