"NLP in India" Meet
7 and 9 January 2008
Hyderabad

A meeting was organized on NLP in India by the NLP Association of India on the occasion of IJCNLP-08 (International Joint conference of Natural Language Processing, 2008). The meeting was open to all NLP researchers in India as well as researchers abroad who wish to contribute to promotion of NLP in India. The following were the agenda items:

1. Research areas to be pursued
2. Linguistic resources
3. Manpower training
4. Increasing the research community
5. Starting journal in NLP

The meeting was organized in two parts. The first part of the meet took place on 7 January 2008 from 5:30 pm -7:30 pm during which all the agenda items were discussed. Three groups were formed on items 2,3 and 5 respectively, to deliberate on the issues and report back in the second part of the meeting on 9 January 2008 from 6:00 pm-8:00 pm. Each group presented a report on its deliberations and an in-depth discussion ensued.

A summary of discussions over the two part meetings follows.

1. Research areas

The following areas in NLP were identified to be important in the Indian context:

- Morphology
- Shallow Parsing
- Parsing
- Named entity recognition
- Transfer grammar
- Language generation
- Linguistic data resources-including annotated data, ontology, WordNet, etc.
- Semantic processing.

Work in these areas will help in applications such as search, information extraction, machine translation, grammar checker, etc.

2. Lexical and Linguistic Data Resources

After an initial discussion the following group was formed on 7 January:

Drs. Dipti M Sharma (IIIT-H), Mallikarjun (CIIL-Mysore),
Pushpak Bhattacharya (IITB), Uma Maheshwar Rao (UoHyd),
Ajay (CDAC, Pune) and Sarmad Hussain (Lahore)

After the group presented its report on 9 January, the following points were agreed to:
- Annotated data preparation is a high priority item.

Cover those major languages for which such data has not been prepared so far. Minority languages can be taken up after that or in parallel.

- Detailed and precise guidelines should be drawn for each annotation task. Common standards across Indian languages should be defined.
- Quality checks are extremely important to make sure that the data is of high quality.
- Raw corpus needs to be collected and made available.
- Parallel corpora are also needed.
- Annotation tasks of high priority are:
  . POS tagging,
  . Chunking
  . Dependency tree bank
  . Sense tagging
- Develop tools and interfaces for effective tagging. People should be able to work from their home.
- Have online test for selecting annotators.
- The following data resources may also be created - Word lists, dictionaries, Wordnet, etc.

3. Manpower Training

The following group was formed:
Drs. Hemant Darbari (CDAC-Pune), Soma Paul (IIIT-H),
Fei Xia (University of Washington).

Here is a summary of discussions:

- A number of degree programmes are running at IIIT-H and other places. Output from these programmes should be hired by the different centres working on NLP. They are highly trained and include people originally from two streams: CS background and Linguistics background. They can constitute the scientific and technological manpower, who can guide others as well as take lead in new initiatives, initially under the guidance of senior researchers.
- Besides the above, there is a need for 6-month certificate training programme or 1-year PG Diploma in which people hired in consortia projects, for example, can be trained.
- CDAC-Pune volunteered to design and run such 6-month or 1-year programmes, which they need in any case for training their own project staff.
- Prof. Fei Xia (University of Washington) promised to provide information regarding four NLP modules which they run at their university:
  . Shallow parsing
  . Parsing
  . Statistical methods
  . Systems and applications
- Prof. Pushpak Bhattacharya cautioned that such programs should be made interesting to people with CS background.
- A 1-month empowerment training was also proposed. It could be run by educational institutions. Need for summer school and winter school was also underlined.

4. Increasing the Research community

Increasing the size of research community was identified as a key task. Existing groups can grow in size and depth. Besides this, new groups should come up. The following were suggested as ways to create new groups:

- Mentor smaller institutions. A major NLP centre should attach to itself a number of interested smaller institutions.
- Run short courses and schools at major centres, so that people can come and gain knowledge. (Many of the initial contacts for mentoring are made at this level.)
- Faculty from smaller institutions may come to do their PhD at the mentor institutions.
- Once a small group is established at a smaller centre, it can be funded for doing a task through DIT.

5. Starting a journal in NLP

The following group was formed:
Drs. Navjyoti Singh (IIIT-H), Rajeev Sangal (IIIT-H),
Vasudeva Varma (IIIT-H).

- It was mentioned that a journal would bring together the research community. It will give visibility to the Indian community as well as help give technical directions.
- It is important to maintain quality of the journal.
- There were suggestions to first start a magazine kind of publication, and later convert it to a journal. It was pointed out that once the publication has a reputation of a magazine, it would never be a scholarly journal.
- There was much discussion on viability of the journal on whether researchers would submit their good quality work to the journal. It was pointed out that most researchers publish their work in conferences today. They can continue to do so, but after that submit a fuller paper to the journal.
- There were fears whether there would be sufficient quantity of research to support a journal.
- It was mentioned that for a journal with two issues per year of 100+ pages each, there needs to be about 15 papers per year. There is sufficient work to produce these many papers.
- It was also mentioned that the authors of best few papers in ICON and other conferences may be encouraged to expand and submit their work to the journal. (Conferences have a very strict limit on number of pages which does not allow researchers to do justice to reporting their work.)
- When it was said that there is no urgency in taking a decision on starting the journal, overwhelming response was to bring out the journal soon. It anyway takes about a year after the planning starts to bring out the first issue.
- There was also much discussion on the name to be used for the journal. One suggestion was to put "Indian Languages" or "South-Asian Languages" in the title, but it was felt to be too restrictive. Names suggested were: Patanjali, Vyakarana etc. Finally a committee was formed to define the scope and the name: Drs. Rajeev Sangal (IIIT-H), Navjyoti Singh (IIIT-H), Pushpak Bhattacharya (IIT-B), Vasudeva Varma (IIIT-H), A Kumaran (Microsoft Research), Dipti M Sharma (IIIT-H) and Aravind Joshi (University of Pennsylvania). The committee should submit its report in 4-weeks.

= Rajeev Sangal

President, NLP Association, India