

NLP in Sri Lanka: Where it is and where it could be

Ruvan Weerasinghe

Language Technology Research Lab
University of Colombo School of Computing

Linguascape

- 3 Major languages (overall literacy >90%)
 - Sinhala (Indo-European) spoken by 70+%
 - Tamil (Dravidian) spoken by 20+%
 - English used as ‘link language’ (<10% literacy)
- Cultures polarized along language divide
 - Ethnic tension since independence (esp. from 1970’s)
 - English the second language!

Linguascape (contd.)

- Very few Sinhala and Tamil language courses at university level
- A single Linguistics programme at a single university
- No shortage of divergent schools of scholars
- Mostly limited to descriptive work

Travails of a 'Returnee'

- Computational Ling-what returns in 1994?
- To a non-existent CL-scape (Sinhala)
 - Several incompatible font encodings
 - Several non-standardized keyboard mappings
 - Undefined collation sequence (despite ancient alphabet!) e.g. discrepancies in dictionaries
 - Virtually no electronic texts (typeset docs)
 - Proprietary software designed for customer lock-in

Travails of a 'Returnee' (contd.)

- Inevitable hibernation
 - Internet needed to be setup (laterally wired)
 - Computer Science needed to be formalized
 - Teaching of any and every CS course
 - Light relief in the form of UG projects
 - Not as 'sexy' as other areas such as graphics!
- Connections only with west (sporadic)
 - Western researchers, languages

Crawling back...

- Introduction of a course in Masters program
- A few Masters projects (not sustainable)
- 2 Sabbatical positions @ INRIA, CMU
- IDRC support (3 years): Pan Asia Network
 - 6 Country programme (ongoing since 2004)
 - Enabler for setting up LTRL
 - With 5 highly enthusiastic RA's (with good first degrees in CS) and a Corpus Linguist!

Crawling back... (contd.)

- PAN L10N deliverables
 - 10m word Sinhala corpus
 - 30k word Sinhala dictionary
 - Commercial grade OCR system
 - Commercial quality TTS system
- Piggy backing projects
 - MPhils in interesting areas (+MAs!)
 - National standards efforts (keyboard, UNICODE fonts, collation sequence), official dictionary

Lessons & Issues

- Excellent people network
 - Publishers (mainly Mac users!)
 - Scholars of all stripes
 - ICT Agency (govt. ICT policy body)
 - Font & software developers
 - Journalists & others
- Up & coming bunch of Computational Linguists

Lessons & Issues (contd)

- Scalability problem
 - No CL programme
 - No short-term training opportunities
 - No conference participation
- Timeframe problem
 - What is the shortest time taken to build a 100m word tagged corpus by a set of untrained folks?
 - It'd be nice to have a Sinhala treebank!
 - ... hey wait a minute – where are the tags?

Possible ways out

- Don't wait for Linguists – use patterns and statistics (let the data 'talk')
 - Adopted for the time being (SMT, word clustering...)
- Start small, keep working on key resources
 - Corpus, tagged/chunked, lexicon/dictionary
- Adopt/adapt successful key technologies
 - Taggers, stemmers, parsers, chunkers,...

Possible ways out (contd)

- Adopt open source/content/community model
 - Exploit the codefest idea
 - Run competitions
 - Setup wiki's
- Forge more active collaborations in Asia
 - Sharing frameworks (lexical, grammar, MT)
 - Exchange training (summer schools/workshops/visiting scholars/professors)
 - Not leaving others to re-invent the wheel (no competition between languages)

After all...

- Access to resources in one's own language
 - a basic human right?
- In this region, it has implications of 'life-and-death' proportions!
 - Mistrust between language groups
 - Poverty cycle for the digitally underprivileged
- Opening up the Language Research space for the sake of the people of Asia...
 - Will the Asian Shuttleworth stand up please?