Visit to Stanford: a Report

Dr. Pushpak Bhattacharyya,
Professor,
Department of Computer Science and Engineering,
IIT Bombay.
pb@cse.iitb.ac.in

This summer (May-July, 2004), I visited Stanford University and other research organizations. Given below is a report of this visit. The description is divided into 7 parts:

A. Visit to Stanford University.
B. One day visit to Microsoft Research, Seattle.
C. One day Visit to the Development Gateway Foundation, World Bank, Washington DC.
D. One day visit to Honda Research Institute, Sunnyvale, California.
E. Titles and Abstracts of talks delivered.
F. Conclusions.
G. The home page of the Stanford Natural Language Processing Group

A. Stanford University

The host institution for the visit was Stanford University. The objective of the visit was to set up collaborative research in Natural Language Processing and related areas between the research groups of IIT Bombay and Stanford University. The facts pertaining to this visit are as under:

1. *Visitor*: Dr. Pushpak Bhattacharyya, Professor, Computer Science and Engineering Department, Indian Institute of Technology, Bombay, India.
2. *Place of Visit*: Computer Science Department, Stanford University.
3. *Duration*: May-July, 2004.
4. *Host and Principal Collaborator*: Prof. Christopher Manning, Computer Science Department, Stanford University.
5. *Research Group*: Natural Language Processing Group of Prof. Manning.
6. *Research group members interacted with (topics in parentheses)*:
    a) Roger Levy ("Probabilistic Parsing", "Psycholinguistic Experiments on Hindi Clauses").
    b) Dan Klein ("Lexicalized and Unlexicalized Parsing").
    c) Kristina Toutanova ("Word Dependencies", "PP Attachement").
    d) Galen Andrews ("Implementation and Installantion of Stanford Parsers").
    e) Teg Grenagar ("Probabilistic Context Free Grammar").
    f) Iddo Lev ("Universal Networking Language", "Logic Puzzles").
    g) Itamar Rosenn ("Logic Puzzles", "Stanford AI courses").
    h) Mona Dieb ("Arabic Wordnet", "Probabilistic Parser for Arabic").
7. *Focus of Research*: Study of the
    a) Use of Statistical Parsers for Semantics Generation from Sentences.
    b) Effect of Placement of Relative Pronouns on the difficulty of Clause Recognition.

8. *Progress in Collaborative Research*:
    a) Experiments have been conducted on Stanford Lexicalized and Unlexicalized parsers to see their performance on different language phenomena in the context of interlingua based MT (work of *Anupama Dutta, Masters student* and *Dr. Rajat Mohanty, Research Linguist* with feedback from *Chris Manning* and *Roger Levy*) .
    b) Hindi corpora have been made accessible in Unicode format (work of *Satish Dethe, Research Associate*). With the availability of Hindi Morph Analyser and POS tagger (work of *Nitin Aggarwal, Masters student, Dr. Bibhuti Mahapatra, Research Linguist* and *Manish Srivastava, PhD student*) in early September, it will be possible for Stanford and IITB NLP members (led by *Roger* Levy) to conduct psycholinguistic experiments on Hindi clause recognition.
9. *Talks delivered*:
    a) Interlingua Based MT and Verb Knowledge Base Creation (abstract at section D.1)
        i. NLP Research Group, CS Department, Stanford.
        ii. Microsoft Research, Seattle (MSR visit report at section B).
        iii. CS Department, University of California Berkeley
        iv. Yahoo, Sunnyvale, California.
    b) Universal Networking Language and Development Gateway Activities (abstract at section D.2)
        i. Development Gateway Foundation, World Bank, Washington DC (DGF visit report at section C).
    c) Wordnet and its Applications (abstract at section D.3)
        i. Natural Language and Knowledge Representation group, CS Department, Stanford.
    d) Can Autonomous Agents Have Natural Language Interfaces?
        i. Honda Research Institute, Mountain view, California.
10. *Course participation*: CS 224N / Ling 237. Natural Language Processing. (Instructor: Prof. Christopher Manning)
11. *Research Group Seminars/Meetings*:
    a) AI (Alternate Mondays).
    b) NLP (Thursdays).
    c) NL and KR (Fridays).
    d) Java-NLP (Tuesdays).
12. *Talks attended* (chronological order; unless otherwise mentioned the speakers were from Stanford University):
    a) "Learning Random Walk Models for Inducing Word Dependency Distributions" by Kristina Toutonova (ICML 2004 paper by Kristina Toutanova, Christopher Manning and Andrew Y. Ng).
    b) "The Stanford NLP Group" by Prof. Christopher Manning.
    c) "Tutorial on Markov Chain Monte Carlo methods" by Ted Grenagar.
    d) "Unlexicalized Learning of Natural Language Structures" by Dan Klein Defense of PhD thesis under the supervision of Prof. Christopher Manning).

e) "Factoring and Mapping- Common Sense Knowledge Representation" by Prof. Ben Kuiper, UT Austin.
f) "Solving Logic Puzzles: From Robust Processing to Precise Semantics" by Iddo Lev (ACL-2004-Workshop-on-Text-Meaning-and-Interpretation paper by Iddo Lev, Bill MacCartney, and Christopher Manning).
g) "Deep dependencies from context-free statistical parsers: correcting the surface Dependency approximation" by Roger Levy (ACL 2004 paper by Roger levy and Christopher Manning).

13. *Research Discussions with other Academicians/Researchers* (topics in parentheses):
    a) Prof. Beth Levin, Linguistics, Stanford ("Verb Hierarchy").
    b) Prof. Paul Kiparsky, Linguistics, Stanford ("Knowledge Base of Sanskrit Words").
    c) Prof. Dan Jurafsky, CS and Linguistics, Stanford ("Statistical and Knowledge Based NLP").
    d) Dr. Byron Dom, Yahoo ("IR and NLP").
    e) Dr. Kentaro Toyoma ("Emerging Markets Project").
    f) Dr. Stephen D. Richardson: Microsoft Research ("Mindnet", "English-Spanish and English-German MT systems").
    g) Dr. Lucy Vanderwende: Microsoft Research {"Statistical MT").
    h) Dr. Arul Menezes: Microsoft Research ("Text Summarization").
    i) Dr. Raman Chandrasekhar ("MT", "IJCAI 2007 in India").
    j) Dr. Srini Narayan, International Computer Science Institute, Berkeley ("Framenet").
    k) Prof. Marti Hearst, Computer Science Department, UC Berkeley ("Knowledge Based Parsing").
    l) Dr. Rakesh Gupta, Honda Research Institute ("Linkage of Open-Mind and Wordnet Knowledge Bases").
    m) Dr. Ambarish Goswami ("Human Mechanical Capability Amplification")

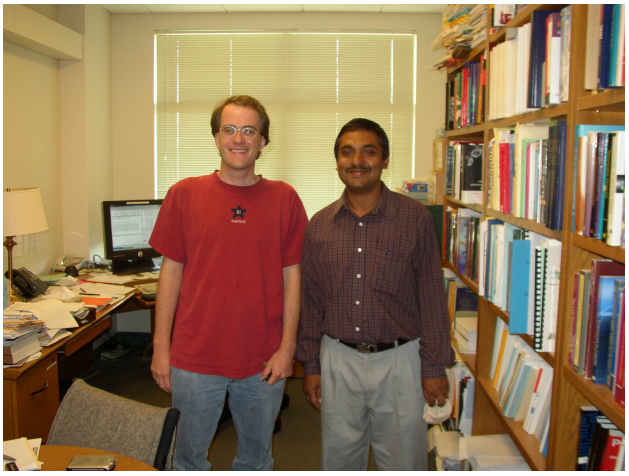

Figure 1 With Prof. Christopher Manning

STANFORD UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

GATES COMPUTER SCIENCE 4A
STANFORD, CA 94305-9040

TEL: (650) 723-0872
TEL: (650) 725-3358
FAX: (650) 725-2588

July 19, 2004

Dear Pushpak,

It was a pleasure having you visit the Stanford NLP (Natural Language Processing) Group this summer. Your visit as a visiting professor at the Stanford Computer Science Department from May through July 2004 was a meaningful one for both you and my NLP Group.

Together we discussed various research topics related to machine translation, statistical parsing and the use and construction of lexical knowledge bases. In addition to our personal discussions, you presented your work at a couple of group meetings, and actively contributed to our various on-going research meetings.

You also interacted with most of my PhD students, in particular, you were able to make available expertise and provide resources to Roger Levy, who was particularly interested in being able to examine data from Hindi for an understanding of psycholinguistic sentence processing.

We appreciated and got considerable value from your visit.

Sincerely,

Christopher Manning
Assistant Professor of Computer Science
Stanford University

Figure 2 Letter from Prof. Manning

B. Microsoft Research, Seattle

I spent a day in Microsoft Research, Seattle, giving a talk and interacting with the NLP group there. The host of the visit was Mr. Kentaro Toyama. Here is a summary of the visit:

1. Dr. Steve Richardson of MSR informed that the Natural Language Processing and Multilingual Computation is the oldest research group of Microsoft dating back to 1991. This was formed at the instance of Mr. Bill Gates himself.
2. The machine translation (MT) group alone consists of about 15 members.
3. MT systems have been developed for English, Spanish, German, French and Chinese. The approach is based on "Logical Forms". I saw the demo of the English to Spanish MT. The domain is technical documentation.
4. I also saw the lexical knowledge network "Mindnet" which has quite a lot of paradigmatic information.
5. In the meeting with Dr. Lucy Vanderwende and Dr. Arul Menezes of MSR, I had further discussions on “Mindnet”. I also saw the MSR summarizer under development.
6. The MT research at MSR- like at many other places- is trying to build systems using parallel corpora and statistical techniques. During my presentation on interlingua based MT, discussions took place on statistical and knowledge based approaches to MT.
7. The lunch meeting was with Mr. Toyama, Dr. Vanderwende, Dr. Menezes, Dr. S. S. Bharat and Mr. S. Chutani. The issue of Microsoft’s University Relations figured amongst miscellaneous discussions.
8. The next generation of MS office package will be augmented with a "search facility", informed Dr. Edward Cutrell of MSR. The s/w ("MS word", e.g.) will have a feature by which one will be able to “information-retrieve” documents and cut and paste matter into the work area. This will work for multiple languages with Unicode representation.
9. I described our work at the Laboratory for Intelligent Internet Research (a set up at the CSE Department, IIT Bombay, funded by Tata Consultancy Services), Center for Indian Language Technology (Funded by the Department of Information Technology, Government of India), Media Lab Asia (funded by the DIT) and the Development Gateway Foundation (A World Bank endeavor funded by DIT).
10. "Emerging Markets" is a very live issue at the MSR- Mr Toyama said. The emerging markets work will hopefully find novel computing solutions for underserved markets, beyond issues of cost. It is expected to apply to all emerging markets worldwide.

C. Development Gateway Foundation (DGF), World Bank, Washington DC

I spent a day at the DGF set-up at the World Bank, Washington DC on 6th July, 2004. The host of the visit was Mr. Jiro Tominaga, Economist, Development Gateway Informatic Program. Here is an excerpt of the proceedings of the day:

1. I made a 2 hour presentation titled "Universal Networking Language and the Development Gateway Activities". The talk was attended by about 10 participants from both technology and business. The technology participants belonged to the

search and content management group and the AIDA (Advanced Information and Data Access) project group.

2. The contents of the presentation were as under:
   Motivation
   Web Access Pattern and the Need for Multilinguality
   Basics of Machine Translation
   UNL Methodology
   Language Divergence
   Hindi Wordnet and Hindi Language Tools
   Verb Knowledge Base
   Multilingual, Meaning Based Search
3. I emphasised that our work on multilinguality in the context of DGF has been and will relate to the Development Gateway Portal whose aim it is to provide a platform for knowledge sharing across countries.
4. The lunch meeting was with Mr. Jiro Tominaga. The discussion centered around the relationship between the country training centers and the country gateways.
5. In the afternoon, I met Mr. Sudhakar Kaveeshwar, Manager, Informatics program, Information Solutions Group along with Mr. Jiro Tominaga to discuss the project planning for the year 2004 on the language technology front. It was decided that we will work on the following problem:

   Problem Statement: Search and Multilingual Access on Topic Page "Meta Data"

   The DGF portal contains pages on specific topics like Population and Reproductive Health, AIDS/HIV, Environment Degradation and so on. (http://topics.developmentgateway.org/alltopics/). These pages are called "Resources" and attached to each resource are a title and an abstract. This title-abstract combination is called the "meta data". The resources are contributed by user(s) of the DGF portal across the world. For each resource, the meta data is created by an editorial team assisted by topic experts. The central motivation for the meta data is to "whet the appetite" of the information requirer. Should the user find the information useful- as suggested by the meta data- he/she will reach the resource. If the user is not comfortable with English, he/she will make an effort to get the resource translated.

   Now, there is a requirement of the meta data being available in multiple languages. Since the size of the meta data is small (at most 10-15 sentences) and since its structure can be controlled, automated methods for multilingual access is feasible.

   It was, therefore, decided that we will convert the meta data into the UNL form and provide search and multilingual generation facility for the meta data.
6. Mr. Mike Pereira, the Content Manager, DGF, underlined the requirement of Searching outside the DGF portal when the user request fails to retrieve the resource (i.e., a set of pages).

D. Honda Research Institute

I had a day's visit at the institute and spoke on Natural Language Processing and NL Interfaces. The host of the visit was Dr. Rakesh Gupta.

a) Dr. Gupta and I exchanged research ideas on possibilities of integrating the OpenMind and the Wordnet knowledge bases.
b) The verb knowledge base being developed at IIT Bombay could enrich such lexical knowledge.
c) In the afternoon I had discussions with Chirag and Vasco (graduate students from UMass, Amherst and CMU respectively) doing there summer internship at HRI.
d) Asimo, the home robot, is one of the famous products of Honda. Dr. Ambarish Goswami of HRI described their very interesting project on "Human Mechanical Capability Amplification" which is meant to augment the muscle power in aged and ailing through implants in the body.

E. Talk Titles and Abstracts

a) *Title*- **Interlingua Based Machine Translation and Verb Knowledge Base Creation**

Abstract- MT systems using interlingua are known to be extremely "knowledge hungry". When the interlingua representation is generated the system needs to commit to the semantics of the sentence. This needs "knowledge content" of the system to be extracted entailing exploitation of high quality lexical knowledge.

While traditional and well known lexical knowledge bases have dealt effectively with nominal concepts, verbal knowledge hierarchies typically have been shallow. Need has always been felt for sophisticated and deeper verbal concept hierarchies which also captures the argument structures of the verbs.

In this talk, we present our work on the creation of a verbal knowledge base that supports the ongoing work on the interlingua based MT. The Universal Networking Language- which is the interlingua representation we adopt- will be introduced, the problems arising from the "language divergences" will be discussed and the verb knowledge hierarchy that is being constructed to solve many of the problems will be described.

b) *Title*- **Universal Networking Language and Development Gateway Activities**

Abstract- In today's world of internationalization and the internet, which has become ubiquitous, the search requirement also has assumed ambitious proportions. One one hand, there is a question answering (QA) kind of demand where the answer to the query is required to be sharp and to the point. On the other hand, demand exists for posing queries in a certain language and retrieving the answer- even if the information is available on the web in a different language.

In this talk we present our work on a system called "AgroExplorer" which is a meaning based, multilingual search engine in the domain of agriculture. The information is stored in a language independent form called the "Universal Networking Language (UNL)". A process called "enconversion" translates the input query into the UNL form. The search returns documents in language independent forms which then are converted into the target language by a process called "deconversion". A focused crawler and a backend enconverter stores the agro-related pages in the UNL form. The indexing is in three levels and complex:

inverted indices are created for semantic subgraphs, concepts and the keywords. The search returns very precise documents which match the meaning content of the query; the process is backed up by concept based search and keyword based search.

c) *Title-* **Wordnet and its Applications**

Abstract- Lexical Knowledge Bases have assumed great importance in today's research and devlopment work natural language processing, machine translation and information extraction. Wordnet is a linked structure of words and concepts and was originally created for English at the Princeton University. Wordnet for other languages soon followed and continue to be developed today.

In this talk- with the flavour of a tutorial- the design principle of the wordnet will first be explained. The lexical and semantic relations will then be described with their types and examples. The applications of the wordnet in three research problems tackled at IIT Bombay, viz., Word Sense Disambiguation, Question Answering and Text summarization will finally be explicated.

d) *Title-* **Can Autonomous agents have Natural Language Interfaces?**

Abstract- If autonomous agents like robots had a natural language based command interface, it would be a very desirable situation to have. However, any natural language based system must solve the classical problems like sense disambiguation, redundancy, co-reference resolution, oblique references and so on. The solution to these problems typically require powerful analyis procedures and large amount of world and lexical knowledge.

In this talk, starting with a brief introduction to natural language processing, we move on to discuss knowledge structures like the WordNet and the OpenMind. We examine the use of these resources in building a language analysis system. We describe our work on analysing English into a semantic net like knowledge representation. A possible use of this meaning extraction could be the initiation of precise actions in autonomous agents.

E. Conclusions

a) Natural Language Processing R & D is huge in the US. The Stanford NLP group, which is amongst the most renowned in the world, is playing a major role in this through its multifarious projects on *Question Answering*, *Statistical Parsing for English, Chinese and Arabic*, *Logic Puzzle Solving*, *Statistical Machine Translation, etc.*

b) Machine Translation between English and Chinese and that between English and Arabic are drawing large corporate funding in the US.

c) The interest in Indian language computing is very significant.

d) Statistical natural language processing is the trend for robust and scalable systems in modern times. However, traditional knowledge based approaches are also receiving renewed attention- mainly a result of the paucity of large training corpora.

e) Language enabled systems (web search, interface to autonomous agents, *etc.*) are very active research areas.

f) Language processing is deemed to play a major role in *reducing the digital divide* .

Home

Index

- **Overview**
- **People**
- **Research**
- **Publications**
- **NLP Links**
- **Local Page**
- **Contact Us**

Projects

- **JavaNLP**
- **Clustering Models**
- **Hypergraph Parsing**
- **Language Learning**
- **Machine Translation**
- **Personalized Pagerank**
- **NL-KR (logic puzzles)**
- **QuASI (ARDA semantics)**
- **Redwoods (Prob. HPSG)**
- **SEER (Edinburgh NER)**

Demos

- **Constrained Clustering**

Downloads

- **Lexicalized Parser**
- **QuASI software**
- **Classifier**

Home

- **NLP**
- **CS Dept**
- **Ling Dept**
- **Stanford**

Stanford NLP logo
(c) Dan Klein 2001

Welcome to the
Stanford Natural Language Processing Group
Homepage

Overview

The Stanford NLP group works on getting computers to process and understand natural human languages (that is, it does *computational linguistics*). The NLP group is part of the Stanford InfoLab.

The group has developed state-of-the-art language technology for doing robust, broad-coverage, scalable natural language processing, including developing the current world's best part-of-speech tagger, a high performance probabilistic parser, stochastic parse selection models for linguistically rich constraint-based lexicalist grammars, and software for mapping from sentences to semantic role representations. General interests include probabilistic models of natural language, grammar induction, word sense disambiguation, deep semantic processing, machine translation, clustering, discriminative probabilistic models, and information extraction. The distingushing feature of our work is effectively combining sophisticated and deep linguistic modeling and data analysis with innovative probabilistic approaches to NLP and machine learning.

People

The NLP group consists of:

- Professors
  - Chris Manning
  - Dan Jurafsky
- Visiting Professors
  - Pushpak Bhattacharyya
- Postdocs
  - Mona Diab
- Ph.D. Students
  - Teg Grenager
  - Sep Kamvar
  - Dan Klein
  - Roger Levy
  - Jeanette Pettibone
  - Kristina Toutanova
  - Bill MacCartney
- MS Students
  - Huy Nguyen
  - Jenny Finkel
  - Galen Andrew
  - Rion Snow
- Alumni/Alumnae
  - Mike Jahr (MS in Computer Science; BS (Hons) in Symbolic Systems)
  - Jonas Kuhn (Visiting scholar; now UTexas)

    - Satoshi Oyama (Visiting Asst Prof. from Kyoto U.)
    - Kristen Parton (BS in Computer Science)
    - Joseph Smarr (MS in Symbolic Systems; now at Plaxo)
    - Cindi Thompson (Visiting Asst Prof. from U. Utah)

Research

Project descriptions:

- Clustering Models
- Hypergraph Parsing
- Language Learning
- Statistical Machine Translation

You can find more details about individual research projects as described by the members of the group.

Publications

- NLP group publications available at the Stanford Database Group publication server
- Also, see the project pages.

Teaching

- NLP courses at Stanford 2003-04
- NLP courses at Stanford 2001-02
- CS224N/Ling237 - Natural Language Processing
    - Final projects 2000
    - Final projects 2001
    - Final projects 2002
    - Final projects 2003
- The NLP Reading Group meets most Friday's at 2:15pm. NLP researchers from local industry are also welcome to attend.

NLP Links

Some other groups at Stanford do some NLP-related research:

- The Computational Semantics Lab at CSLI
- The LinGO/LKB project at CSLI
- The Stanford Edinburgh Entity Recognition (SEER) Project
- The Stanford Speech Processing Group
- Martin Kay

Contact Information

We're interested in new students (once admitted to Stanford), new projects, and opportunities for employing NLP.
Information about the group? Feel free to email Chris.
Comments about the web page? Feel free to email Jenny.