CV of Anil Kumar Singh

Independent Researcher and Translator

nlprnd@gmail.com

EDUCATION

- Ph.D. (Computational Linguistics), 2010
  International Institute for Information Technology, Hyderabad, A.P., India
- B.E. (Mech. Engg.), 1990
  NIT, Jaipur, Rajasthan, India

LANGUAGES KNOWN

- English (Spoken: Fluent, Written: Near native)
- Hindi (Spoken: Native, Written: Native)

LANGUAGES FAMILIAR WITH

- French (Spoken: A little, Written: Significantly better than spoken, especially spoken language comprehension)
- Braj (Spoken: Moderate, Written: Moderate)
- Awadhi (Spoken: Moderate, Written: Moderate)
- Rajasthani (Spoken: Moderate, Written: Moderate)
- Punjabi (Spoken: Moderate, Written: Very little)
- Sanskrit (Studied till high school, but mostly forgotten)

WORK EXPERIENCE

- **Post-doctoral Researcher, TLP, LIMSI-CNRS, Paris, France**. (May, 2012 to to April, 2013)
- **Senior Assistant Professor, School of Computer Engg., KIIT University, Bhubaneswar, India**. (May, 2011 to April, 2012)
- **Research Scientist, LTRC, IIIT, Hyderabad**. (July, 2009 to May, 2011)
- **Consultant, OutBox Edutainment Pvt. Ltd., Hyderabad**. (September, 2008 to January, 2011)
- **Research Assistant, LTRC, IIIT, Hyderabad**. (July, 2003 to July, 2009)

- Prior to joining PhD, worked as software developer, freelance writer, editor, translator, part-time teacher/lecturer etc., mostly in New Delhi. Continued translating even after joining PhD in 2003, but it was mostly voluntary and unpaid, though it was of the highest quality of writing, including poetry.

CAT TOOLS

I can use any available tools. I have tried some of them, such as gTranslate, Podit and OmegaT. The problem is that hardly any of them has good resources (dictionaries, MT systems) available with them for the English-Hindi pair, although they can be used for translation even without these resources, their utility is not as much as for languages like French and Spanish.

I have the skills to build or enhance CAT tools myself.

TRANSLATION INTERESTS/EXPERTISE

- Arts and Humanities
- Science and Technology.
- Computers and related topics
- Copywriting
- General
- News, Current Affairs and Politics
- History
- Literature / Poetry
- Localization
- Translation ***for*** Machine Translation and Natural Language Processing tasks, such as for developing a Machine Translation system

ADDITIONAL INFORMATION FOR TRANSATION JOBS

I am currently not doing any full time job, so I can work on most days. However, I can only work part-time.

I will prefer translating creative writing or political/academic/journalistic writing, but I can translate other kinds of text too.

I have a long experience of translating between the pair English and Hindi, and have translated various kinds of writing over the last twenty year (though not regularly). Whatever kind of writing I get, if I accept to translate it, I am very confident of being able to provide very good translations. I have translated some very well known authors.

As I am, professionally, a researcher in Computational Linguistics and have experience in research and development on Machine Translation, I can also work on tasks which require such skills. I have a long experience in programming.

I have also been a technical writer, an academic writer and a literary writer: in English as well as Hindi.

As a computational linguist and Machine Translation researcher, I have some familiarity with translation theory.

Hindi is my native language. In English, I have near native writing skills.

RESEARCH INTERESTS

My research interests can be summarized as (not necessarily in that order):

- Machine Translation (MT), especially Statistical MT
- Creation, processing and use of language resources (including ontologies)
- Use of machine learning and statistical techniques for solving language processing problems
- Modeling and processing of events and temporal information in Natural Languages (NLs) for Information Extraction
- Information Retrieval and Crosslingual Information Retrieval
- Processing of noisy and unstructured text
- Linguistic similarity in the broadest sense
- Writing systems and computation
- Historical linguistics and the use of computational techniques for phylogenetic study of languages
- Computational phonology and morphology
- Speech recognition and synthesis
- Natural Language Engineering, including development of NL software

SOME RECENT R&D CONTRIBUTIONS AND INNOVATIONS

- Initiated and developed a tool called Questimate[1] for Machine Translation quality estimation that includes a feature extraction API, Weka based prediction and classification system, as well as a GUI for editing features and for analysis of their relationships with the help of statistical plots.
- Developed a tool called LatticeFst[2] for computation of n-gram posterior probabilities from word or phrase lattices (for Machine Translation purposes).
- Participated in a study of Machine Translation quality estimation, focusing on how the problem should be formulated and evaluated, and the associated problems.

---

[1] https://bitbucket.org/anilsingh/questimate

[2] https://bitbucket.org/anilsingh/latticefst

- Initiated and developed an open source platform called Sanchay[3] for Natural Language Processing. It is primarily meant for researchers, but many facilities and tools in it can be useful for common users too. It is still a work-in-progress.
- An important part of Sanchay is a workbench for automatic annotation using various Machine Learning (ML) libraries. The idea is to allow a user (a researcher) to use a simple GUI to select features, select the ML techniques, load annotated data directly and perform training for a specific purpose. The trained tool then can be used for automatic annotation. The main goal was to make the task of performing experiments on language data using ML techniques easier for those who are not experts.
- Other parts of Sanchay include annotation interfaces (syntactic annotation, parallel corpus alignment, PropBank annotation etc.). The Sanchay Syntactic Annotation interface is being used (especially) for Indian languages in many research centres.
- Designed a new query language for corpora represented as ‘tangled trees’ or ‘threaded trees’ (trees with additional links between nodes across branches) that can encode multiple levels of annotation in the same file that can be stored in simple XML format. The language has a very simple, intuitive and concise syntax and high expressive power. It allows not only to search for complicated patterns with very short queries but also allows manipulation of the data and specification of arbitrary return values. The latter can be used to extract a wide variety of features from the data for any purpose, e.g. for Machine Learning techniques. The conditions for the queries can be combined with AND and OR operators and they can be nested. The language also allows multiple sources and destinations to be specified in a single query.
- A study of linguistic similarity, including proposing a generalized definition, a typology of linguistic similarity and innovations in techniques for estimating different kinds of similarities.
- Major contributions to the development of treebanks for Indian languages
- Extended the TimeML (Time Markup Language) to better take into account information related to tense, aspect and modality
- Proposed a Computational Phonetic Model of Scripts (CPMS) that has been useful in improving the results for many NLP applications. Several publications related to this have been published.
- Proposed an Architectural View of the Typology of Writing System that can better inform the development of techniques for text processing.
- Proposed several new similarity and association measures, some of which are based on the CPMS (e.g. Cognate Coverage Distance and Phonetic Distance of Cognates) while others are purely statistical (e.g. Normalized Conditional Mutual Information).
- Have been involved in the language resource development activities of the centre ever since I joined the current institute (IIIT, Hyderabad, India).
- Have contributed to Machine Translation projects (past and present) in the centre.

---

[3] http://sanchay.co.in

- Played a primary role in developing offline and online word games for Indian languages for Outbox Edutainment Pvt. Ltd. (joro.in), which were the first for these languages.
- Worked on several other NLP/CL problems with results that could be published.

COURSES FINISHED DURING PHD

- Computational Linguistics I
- Natural Language Processing
- Computational Linguistics II
- Natural Language Databases (Project based course on designing NL based dialog systems for querying databases)
- Designing Speech Systems
- Robotics (project based course on designing an NL interface for UMR robots)
- Advanced Problem Solving
- Machine Learning
- Linguistic I
- Linguistic II

SUBJECTS CLEARED FOR BREADTH QUALIFIER EXAM

- Data Structures and Algorithms
- Database Management Systems

SOME R&D RELATED SKILLS

1. **Machine Learning**: Clustering, CRF, SVM, Maximum Entropy and a little about some other learning approaches
2. **Statistical NLP tools**: Moses, GIZA++, SRILM, OpenNLP, Weka etc. (I integrated Java libraries for some of these in Sanchay)
3. **Programming**: Java, Perl, Python, C, C++ etc.
4. **Databases**: MySql, Oracle, Derby
5. **Operating Systems**: Linux, Windows
6. **IDEs**: Netbeans, Eclipse, Visual Studio etc.
7. **Others**: Numerous libraries, tools etc. used for research

COURSES TAUGHT

- Natural Language Processing, Spring, 2012

- Computer Systems and Programming (Labs), Spring, 2012
- Computer Systems and Programming (Lectures), Autumn, 2011
- Computer Systems and Programming (Labs), Autumn, 2011
- Artificial Intelligence, Autumn, 2011
- Computer Environment and Scripting & IT Workshop 1A, Autumn, 2009
- Language Typology and Universals, Autumn, 2007-08 (A few lectures)
- Computational Linguistics I, Autumn, 2007-08
- Computer and Scripting I, Autumn, 2006-07 (A few lectures)
- Computer and Scripting II, Spring, 2005-06

EVENTS ORGANIZED

- An Orientation-cum-Training Program on NLP at the KIIT University, Bhubaneswar, 2012 (Co-organizer: Pramod Rout, CIIL, Mysore, India)
- An Introductory Workshop on Natural Language Processing at KIIT University, Bhubaneswar, 2011 (Co-organizer: Sriram Chaudhury, KIIT University, Bhubaneswar, India)
- IJCNLP 2008 Workshop on NLP for Less Privileged Languages, Hyderabad, India
- IJCNLP 2008 Workshop on Named Entities Recognition (NER) for South and South East Asian Languages, Hyderabad, India (Co-organizers: Rajeev Sangal and Dipti Misra Sharma, IIIT-Hyderabad, India)
- NLPAI Machine Learning Contest 2007 on Named Entities Recognition for South Asian Languages, Hyderabad, India (Co-organizers: Rajeev Sangal and Dipti Misra Sharma, IIIT-Hyderabad, India)

GUEST LECTURES

- Lectures on NLP as part of an orientation-cum-training program on NLP at the University of Kashmir, Srinagar, November, 2011.
- Lectures on NLP as part of an orientation-cum-training program on NLP at the Banaras Hindu University (BHU) Varanasi, January, 2012.
- Lectures on NLP as part of an orientation-cum-training program on NLP at the Lucknow University, Lucknow, February, 2012.
- Lectures on NLP as part of an orientation-cum-training program on NLP at the Jadavpur University, Kolkata, February, 2012.
- Lectures on NLP as part of an orientation-cum-training program on NLP at the KIIT University, Bhubaneswar, February, 2012.

THESES SUPERVISED

(Being a senior researcher in the research centre, these theses were informally co-guided. Published papers related to these were co-authored together with these students. The names of the supervisors are given below.)

1. Building a Weakly Supervised Dependency Parser and Language Identification in Multilingual Documents. Jagadeesh Gorla. M.S. by Research Thesis. Guided by Prof. Rajeev Sangal. LTRC, IIIT, Hyderabad. 2009.
2. A Java Implementation of an Extended Word Alignment Algorithm Based on the IBM Models. Chinnappa Guggila. M. Tech. Thesis. Guided by Prof. Dipti Misra Sharma. LTRC, IIIT, Hyderabad. 2009.
3. Effective Query Translation Techniques for Cross-Language Information Retrieval. Sethuramalingam Subramaniam. M.S. by Research Thesis. Guided by Prof. Vadudeva Verma. LTRC, IIIT, Hyderabad. 2009.
4. Cognate Identification and Phylogeny in Dravidian Languages. Taraka Rama. M. Tech. Thesis. Guided by Prof. Lakshmi Bai. LTRC, IIIT, Hyderabad. 2009.
5. Improving the Performance of the Link Parser. Y. Viswanath Naidu. M. Phil. Thesis. Guided by Prof. Dipti Misra Sharma. LTRC, IIIT, Hyderabad.

OTHER ACTIVITIES AND INTERESTS

- Writing poems, essays etc. and (unpaid/voluntary) translation the works of some major writers into Hindi
- Photography
- Music, paintings, movies, etc. (mainly consumption, not creation)

RESEARCH PUBLICATIONS

- *LIMSI Submission for the WMT'13 Quality Estimation Task: an Experiment with n-gram Posteriors.* G. Wisniewski, A. K. Singh and F. Yvon. In Proceeding of the ACL 2013 Workshop on Machine Translation (WMT'13). Sophia, Bulgaria. August, 2013.
- *Design and Analysis of a Large Corpus of Post-Edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-Edition.* G. Wisniewski, A. K. Singh, Natalia Segal and F. Yvon. In Proceeding of the Machine Translation Summit XIV. Nice, France. September, 2013.
- *Quality Estimation for Machine Translation: Some Lessons Learned*. G. Wisniewski, A. K. Singh and F. Yvon. Machine Translation (Special Issue on Quality Estimation), 2013.
- *A Concise Query Language with Search and Transform Operations for Corpora with Multiple Levels of Annotation.* Anil Kumar Singh. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC). Instanbul. 2012.
- *Part-of-Speech Annotation with Sanchay.* Anil Kumar Singh. In Proceedings of the National Seminar On POS Annotation for Indian Languages: Issues & Perspectives.

Mysore, India. 2011.

- *Challenges and Opportunities in Automatically Building Bilingual Lexicon from Web Corpus*. Kiran Pala, Anil Kumar Singh and S. V. Gangashetty. In Proceedings of the Seminar on Linguistic and Language Development In Jammu and Kashmir with Special Reference to Tribal Languages of the State. Srinagar, Kashmir. 2011.
- *Games for Academic Vocabulary Learning Through a Virtual Environment.* Kiran Pala, Anil Kumar Singh and S. V. Gangashetty. In Proceedings of the International Conference on Asian Language Processing. Penang, Malaysia. 2011.
- *First-of-Their-Kind "Total" Telugu Word Games and a User-Friendly Solution for Telugu Text Input in Digital Devices.* Hareesh Viriyala and Anil Kumar Singh. International Telugu Internet Conference. Silicon Valley, California, US. 2011.
- *Transliteration as Alignment vs. Transliteration as Generation for the Purpose of Crosslingual Information Retrieval.* Anil Kumar Singh, Sethuramalingam Subramaniam and Taraka Rama. Traitement Automatique des Langues, Special Issue on Multilingualism and NLP. Vol. 51, Number 2. 2010.
- An Integrated Digital Tool for Accessing Language Resources. Anil Kumar Singh and Bharat Ambati. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC). Malta. 2010.
- Frame Extraction and Verb Classification from Treebank for Hindi and Telugu. Sudheer Kolachina, Prasanth Kolachina, Anil Kumar Singh, Viswanath Naidu and Samar Husain. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC). Malta. 2010.
- Improving the Performance of the Link Parser. Y. Viswanatha Naidu, Anil Kumar Singh, Dipti Misra Sharma and Akshar Bharati. In Proceedings of the International Conference on Asian Language Processing. Singapore. 2009.
- From Bag of Languages to Family Trees From Noisy Corpus. Taraka Rama and Anil Kumar Singh. In Proceedings of the Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria. 2009.
- Experiments in CLIR Using Fuzzy String Search Based on Surface Similarity. Sethuramalingam S, Anil Kumar Singh and Pradeep Dasigi. In Proceedings of the 32nd Annual SIGIR Conference. Boston, Massachusetts. 2009.
- Modeling Letter to Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training. Taraka Rama, Anil Kumar Singh and Sudheer Kolachina. In Proceedings of the NAACL Student Research Workshop. Boulder, Colorado. 2009.
- Review of 'Translation Equivalence, An Essay in Theoretical Linguistics', M. K. C. Uwajeh, Lincom GmbH. Anil Kumar Singh. Linguist List. 2008.
- A Graph Based Method for Building Multilingual Weakly Supervised Dependency Parsers. Jagadeesh Gorla, Anil Kumar Singh, Rajeev Sangal, Karthik Gali, Samar Husain and Sriram Venkatapathy. In Proceedings of the 6th International Conference on Natural Language Processing (GoTAL). Gothenburg, Sweden. 2008.
- Estimating the Cost of Adapting the Resources of One Language for Another. Anil Kumar Singh, Kiran Pala and Harshit Surana. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco. 2008.

- Open Source Software and Object Oriented Technology. Anil Kumar Singh. Tutorial at the Conference on Free and Open Source Software (FOSSCONF). Chennai, India. 2008.
- An Outline of a Multilingual Natural Language Text and Speech Interface for Computing Devices in the South Asian Context. Anil Kumar Singh. In Proceedings of the IUI Workshop on Intelligent User Interfaces for Developing Regions (IUI4DR). Canary Islands, Spain. 2008.
- Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. Anil Kumar Singh (Ed.). Hyderabad, India. Asian Federation of Natural Language Processing. 2008.
- Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages. Rajeev Sangal, Dipti Misra Sharma and Anil Kumar Singh (Ed.). Hyderabad, India. Asian Federation of Natural Language Processing. 2008.
- Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going? Anil Kumar Singh. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages. Hyderabad, India. 2008.
- Named Entity Recognition for South and South East Asian Languages: Taking Stock. Anil Kumar Singh. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages. Hyderabad, India. 2008.
- A Mechanism to Provide Language-Encoding Support and an NLP Friendly Editor. Anil Kumar Singh. In Proceedings of the Third International Joint Conference on Natural Language Processing. Hyderabad, India. 2008.
- A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages. Harshit Surana and Anil Kumar Singh. In Proceedings of the Third International Joint Conference on Natural Language Processing. Hyderabad, India. 2008.
- Multilingual Akshar Based Transducer for South and South East Asian Languages which Use Indic Scripts. Anil Kumar Singh and Harshit Surana. In Proceedings of the Seventh International Symposium on Natural Language Processing. Pattaya, Thailand. 2007.
- Grammar and Multilingualism. Kiran Pala and Anil Kumar Singh. At the National Seminar on the Emerging Linguistic Scene in North East India. Shillong, India. 2007.
- Identification of Languages and Encodings in a Multilingual Document. Anil Kumar Singh and Jagadeesh Gorla. In Proceedings of the 3rd ACL SIGWAC Workshop on Web As Corpus. Louvain-la-Neuve, Belgium. 2007.
- More Accurate Fuzzy Text Search for Languages Using Abugida Scripts. Anil Kumar Singh, Harshit Surana and Karthik Gali. In Proceedings of ACM SIGIR Workshop on Improving Web Retrieval for Non-English Queries. Amsterdam, Netherlands. 2007.
- Disambiguating Tense, Aspect and Modality Markers for Correcting Machine Translation Errors. Anil Kumar Singh, Samar Husain, Harshit Surana, Jagadeesh Gorla, Chinnappa Guggilla and Dipti Misra Sharma. In Proceedings of the Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria. 2007.
- Exploring Translation Similarities for Building a Better Sentence Aligner. Anil Kumar Singh and Samar Husain. In Proceedings of the 3rd Indian International Conference on Artificial Intelligence. Pune, India. 2007.
- Extraction and Translation of Multi-Word Number Expressions. Anil Kumar Singh. In Proceedings of the 3rd Indian International Conference on Artificial Intelligence. Pune,

India. 2007.

- A Java Implementation of an Extended Word Alignment Algorithm Based on the IBM Models. Chinnappa Guggilla and Anil Kumar Singh. In Proceedings of the 3rd Indian International Conference on Artificial Intelligence. Pune, India. 2007.
- Using a Single Framework for Computational Modeling of Linguistic Similarity for Solving Many NLP Problems. Anil Kumar Singh and Harshit Surana. In Proceedings of Eurolan Doctoral Consortium. Iasi, Romania. 2007.
- Can Corpus Based Measures be Used for Comparative Study of Languages? Anil Kumar Singh and Harshit Surana. In Proceedings of the Ninth Meeting of ACL Special Interest Group on Computational Phonology and Morphology. Prague, Czech Republic. 2007.
- Study of Cognates among South Asian Languages for the Purpose of Building Lexical Resources. Anil Kumar Singh and Harshit Surana. Journal of Language Technology. Dept. of IT, India. 2007.
- Using a Model of Scripts for Shallow Morphological Analysis Given an Unannotated Corpus. Anil Kumar Singh and Harshit Surana. ADD-2 Workshop on Morpho-Syntactic Analysis. Bangkok, Thailand. March, 2007.
- A Framework for Computational Processing of Spelling Variation. Anil Kumar Singh. Conference on New Ways of Analyzing Variation (NWAV-35). Columbus, Ohio. November, 2006.
- A Computational Phonetic Model for Indian Language Scripts. Anil Kumar Singh. In Proceedings of the Fifth International Workshop on Writing Systems. Nijmegen, The Netherlands, October, 2006.
- Study of Some Distance Measures for Language and Encoding Identification. Anil Kumar Singh. In Proceeding of ACL 2006 Workshop on Linguistic Distances. Sydney, Australia. July 2006.
- Building An Integrated Digital Tool for Language Resources. Anil Kumar Singh. Issue statement for the Digital Tools Summit 2006. East Lansing, Michigan. June 2006.
- Review of 'Language of Time', (Ed.) Inderjeet Mani, James Pustejovsky, and Robert Gaizauskas, Oxford University Press. Anil Kumar Singh. Linguist List. 2005.
- Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs. Anil Kumar Singh and Samar Husain. In Proceedings of ACL 2005 Workshop on Parallel Text. Ann Arbor, Michigan. June 2005.