

## SANDIPAN DANDAPAT

VB 304A, Postgraduate Residences  
Campus Residences  
Ballymun Road , Dublin 9  
IRELAND

Email: [sandipandandapat@gmail.com](mailto:sandipandandapat@gmail.com)

Cell: +353 860326273

Homepage: <http://www.computing.dcu.ie/~sandipan/>

## INTERESTS

---

**Natural Language Processing, Computational Linguistics, Machine Learning and Artificial Intelligence**

## EDUCATION

---

**Ph.D.** Centre for Next Generation Localization, School of Computing, Dublin City University (from Mar. 2009)  
*Advisors:* Prof. Andy Way and Dr. Sara Morrissey

**M.S.** Department of Computer Science & Engineering, IIT Kharagpur, INDIA (Jan. 2005 – Feb. 2008)

*Advisors:* Prof Sudeshna Sarkar and Prof Anupam Basu, CSE, IIT Kharagpur

*Thesis Title:* Part-of-Speech Tagging for Indian Languages

*Description:* The aim of the work is to develop a Part-of-Speech tagger for Indian Languages especially for resource poor languages. The key techniques used are different machine learning approaches (especially graphical models like HMM, MaxEnt and CRF). The system is being tested on Indian languages (Hindi and Bengali). The system can be used in a variety of applications like search engines, Machine Translation systems, etc.

*Courses Taken:* Machine Learning, Intelligent Systems, Data Mining, Human Computer Interaction.

*CGPA:* 9.51/10

**Post Graduate Diploma (Computational Linguistics).** IIIT Hyderabad, INDIA (2002 – 2004)

*CGPA:* 7.72/10

**B.Tech.** Computer Science & Engineering, Haldia Institute of Technology, INDIA (1998 – 2002)

*Percentage:* 73.3%

**Higher Secondary (XII).** West Bengal Council of Higher Secondary Education. 1997.

*Percentage:* 67.2%

**Madhyamik (X).** West Bengal Board of Secondary Education, 1995.

*Percentage:* 79.2%

## EXPERIENCES

---

**Research Vendor, Microsoft Research Lab India Private Limited,** (from July 2008 – 13<sup>th</sup> January, 2009)

I have worked on methods for fast creation of human annotated data (which is crucial for supervised machine learning algorithms) for several syntactic processing tasks. The task mainly involves creation of user-friendly tools for complex annotation task and incorporating machine intelligence.

**Research Intern, Microsoft Research Lab India Private Limited,** (from March 2007 – February 2008):

I am working on developing tools and algorithms for syntactic processing of Indian language texts.

**Senior Project Officer, Communication Empowerment Laboratory, IIT Kharagpur** (from March 2007 – February 2008):

Sponsored by Department of Information Technology, Government of Information Technology, I have been working on the project IL-IL-MT – Development of Indian Language (IL) to Indian Language (IL) Machine Translation System (MT). My responsibility as a senior project officer is to coordinate with team members to ensure timely roll-out of the tools and design/development of few important framework/modules for different tasks.

**Junior Project Officer, Communication Empowerment Laboratory, IIT Kharagpur** (from July 2004 – February 2007):

Sponsored by Media Lab Asia, I have been working on the project *Sanyog* – Multimodal Communication Tool for Children with Cerebral Palsy since July 2004. My responsibilities as a junior project officer include basic research in Indian language NLP and development of related tools.

**Research Assistant, Language Technology Research Centre, IIIT Hyderabad** (July 2002 – February 2004):

- October 2002 – February 2004
  - Experimental English-Hindi Machine Translation System, Shakti
- December 2002 – August 2003
  - Merging Lexical Resources

**B.Tech Project** (December 2001 – June 2002):

*Title:* Mood Identification of Human Faces

*Advisor:* Mr. Biswapati Jana, CSE, Haldia Institute of Technology

*Summary:* C and VB program which takes a facial image and identifies the mood. The program prepares a knowledge base looking at different facial expressions and compares the new image with the already existing expressions in the knowledge base. The program consists of edge detection, filtering, clustering and knowledge extraction.

**Student Intern, Jadavpur University, INDIA** (May 2001 – August 2001):

Taking a natural language query, the SQL query is generated and executed on databases. The output is in the same language as of the query language. The language of the database is different from the query language. This is basically a NLIR system. This project was developed in VC++ using MFC.

---

## COMPUTER SKILLS

**Programming Languages & Softwares:** Perl, C, C#, SQL, UNIX Shell Programming

**Platforms:** GNU Linux

---

## PUBLICATIONS, AWARD

### Publications

- **Sandipan Dandapat**, Sara Morriessy, Sudip Kumar Naskar and Harold Somers. 2010. Statistically Motivated Example-based Machine Translation using Translation Memory. In *Proceedings of the 8th International Conference on Natural Language Processing, ICON 2010*, Kharagpur, India. pp. 168-177.
- **Sandipan Dandapat**, Sara Morriessy, Sudip Kumar Naskar and Harold Somers. 2010. Mitigating Problems in Analogy-based EBMT with SMT and vice versa: a Case Study with Named Entity Transliteration. In *Proceedings of the 24th Pacific Asia Conference on Language Information and Computation, PACLIC 2010*, Sendai, Japan.
- **Sandipan Dandapat**, Mikel L. Forcada, Declan Groves, Sergio Penkale, John Tinsley and Andy Way. 2010. OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System. In *Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010*, Reykjavik, Iceland. pp. 121-126.

- Sergio Penkale, Rejwanul Haque, **Sandipan Dandapat**, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, Andy Way. 2010. MaTrEx: The DCU MT System for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, ACL 2010*, Uppsala, Sweden. pp. 143-148.
- Sara Morrissey, Harold Somers, Robert Smith, Shane Gilchrist and **Sandipan Dandapat**. Building a Sign Language corpus for use in Machine Translation. In *Proceedings of the 4th Workshop on Representation and Processing of Sign Languages: Corpora for Sign Language Technologies*, 2010. Valetta, Malta. pp. 172-177.
- Harold Somers, **Sandipan Dandapat** and Sudip Kumar Naskar. A review of EBMT using proportional analogy. In *Proceedings of the 3rd Workshop on Example-Based Machine Translation*, 2009. Dublin, Ireland. pp. 53-60.
- **Sandipan Dandapat**, Priyanka Biswas, Monojit Choudhury and Kalika Bali. Complex Linguistic Annotation - No Easy Way Out! A Case from Bangla and Hindi POS Labeling Task. In *Proceedings of the 3rd Linguistic Annotation Workshop (LAW - III) 2009, a Workshop at the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore. pp. 10-18.
- Rejwanul Haque, **Sandipan Dandapat**, Ankit Kumar Srivastava, Sudip Kumar Naskar and Andy Way. English-Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009. In *Proceedings of the Named Entities Workshop (NEWS) 2009, a Workshop at the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore. pp. 104-107.
- Cohan Sujay Carlos, Monojit Choudhury, **Sandipan Dandapat**. Large-Coverage Root Lexicon Extraction for Hindi. In *Proceedings of the 12th Conference of the European Chapter of the ACL, EACL 2009*. Athens, Greece. pp. 121-129.
- Priyanka Biswas, **Sandipan Dandapat**, Kalika Bali, Monojit Choudhury. A Corpus-based Study of ক্রে (kare) in Bangla: Theoretical and Computational Perspectives. In *Proceedings of the 6th International Conference on Natural Language Processing, ICON 2008*. Pune, India. pp.
- Sujay Kumar Saha, Sanjay Chatterjee, **Sandipan Dandapat**, Sudeshna Sarkar and Pabitra Mitra . A Hybrid Approach for Named Entity Recognition in Indian Languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, NERSEAL 08*. Hyderabad, India. <http://lrc.iit.ac.in/ner-ssea-08/>. pp. 17-24.
- Tirthankar Dasgupta, **Sandipan Dandapat**, Anupam Basu. A Prototype Machine Translation System from Text-to-Indian Sign Language. In *Proceedings of the IJCNLP-08 Workshop NLP for Less Privileged Languages, NLPLPL 08*. Hyderabad, India. <http://lrc.iit.ac.in/nlp-lpl-08/>. pp. 19-26.
- Debasis Mandal, **Sandipan Dandapat**, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar. Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources. In *Proceedings of 8th Workshop of the Cross-Language Evaluation Forum*, Budapest, Hungary, 19-21 Sept 2007, LNCS, Springer.
- **Sandipan Dandapat**, Sudeshna Sarkar and Anupam Basu. Automatic Part-of-Speech Tagging for Bengali: An approach for Morphologically Rich Languages in a Poor Resource Scenario. In *Proceedings of the Association of Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp. 221-224.
- **Sandipan Dandapat**. Part-of-Speech Tagging and Chunking with Maximum Entropy Model. In *Proceedings of the SPSAL Workshop*, IJCAI.2007.
- **Sandipan Dandapat** and Sudeshna Sarkar. Part of Speech Tagging for Bengali with Hidden Markov Model. In *Proceedings of the NLP AI Machine Learning Contest*, [http://lrc.iit.ac.in/nlpai\\_contest06](http://lrc.iit.ac.in/nlpai_contest06). June 2006.
- **Sandipan Dandapat**, Pabitra Mitra and Sudeshna Sarkar. Statistical Investigation of Bengali Noun-Verb (N-V) Collocations as Multi-word Expressions. In *Proceedings of the Symposium on Modeling and Shallow Parsing of Indian Languages (MSPIL 2006)*, IIT Bombay, INDIA, pp. 230-233, April 2006.
- **Sandipan Dandapat**, Sudeshna Sarkar and Anupam Basu. A hybrid part-of-speech tagging and its application to Bengali. In *Proceedings of the International Conference on Computational Intelligence (ICCI 2004)*, Istanbul, TURKEY, pp. 169-172.

- Vibhab Agrwal, **Sandipan Dandapat**, Dipti Mishra Sharma and Rajeev Sangal. Linking Monolingual Resource with Bilingual Resource. *Proc. of the Symposium on Indian Morphology Phonology and Language Engineering (SIMPLE 2004)*. IIT Kharagpur, INDIA, March 2004

**Award**

- Awarded first prize in NLP AI Machine Learning Contest 2006 for Part of Speech Tagging and Chunking for Indian Languages.

## **REFERENCES**

---

**Prof Andy Way**

CNGL, School of Computing  
Dublin City University, Dublin 9, IRELAND

**Dr. A. Kumaran**

Microsoft Research Lab India Pvt. Ltd,  
Bangalore – 560080, INDIA

**Prof Anupam Basu**

Department of Computer Science & Engineering,  
IIT Kharagpur, West Bengal, INDIA -721302

## **PERSONAL DETAILS**

---

**Gender:** Male

**Date of birth:** 1<sup>st</sup> February, 1980

**Present Citizenship:** Indian