Arun **Jayapal**

MACHINE LEARNING · NATURAL LANGUAGE PROCESSING · DATA MINING

Plot no. 61, Mount view residency, Opp. Spring Dale Layout, Karapalli, Onnalvadi Post Hosur, Tamil Nadu, INDIA 635 109

☎ (+91) 984-230-9803 | ✉ jayapala@tcd.ie | ⌂ https://arunjeyapal.github.io/ | arunjeyapal | arunjeyapal

"The more I learn, the more I know of the unknown"

Education

Trinity College Dublin (University of Dublin)
Dublin, IRELAND

PHD. IN COMPUTER SCIENCE (COMPUTATIONAL LINGUISTICS) — *Sep. 2013 - Apr. 2018*

- Scholarship covering fees and stipend funded by CNGL-II (now ADAPT centre) from Science Foundation Ireland (SFI)

The University of Sheffield
Sheffield, UNITED KINGDOM

MSC. IN COMPUTER SCIENCE WITH SPEECH AND LANGUAGE PROCESSING — *Sep. 2010 - Aug. 2011*

- Graduated with 2:1 grade (First-class)

Anna University(Arunai Engineering College)
Chennai, INDIA

B.E. IN COMPUTER SCIENCE AND ENGINEERING — *Jun. 2003 - May. 2006*

- Graduated with First-class

Government Polytechnic College (Directorate of Technical Education)
Krishnagiri, Tamil Nadu, INDIA

DIPLOMA IN COMPUTER SCIENCE AND ENGINEERING — *Jun. 2000 - May. 2003*

- Graduated with First-class

PhD thesis

Finding diachronic sense changes by Unsupervised methods A probabilistic model (generative model) has been proposed to identify the time of the emerging sense in the time-varying sense disambiguation problem. Further I have used Expectation Maximization and Gibbs sampling procedures to get the parameter estimates of the proposed probabilistic model.

Skills

Programming	C++, JAVA, Python (2.7 & 3.5+), R, MATLAB, LaTeX
NLP/ML tools	NLTK, Thrax, scikit learn, pytorch, stanford coreNLP, etc.,
Databases	PostGRE Sql
Worked on	Named entity recognition (used state-of-the art classifiers HMM and CRF as part of Master's thesis work), co-reference resolution (as a part of master's thesis, I built a rule-based classifier), sentiment analysis (text classification for that matter) using different state-of-the-art classifiers, Bayesian modelling (constructed own model and used Expectation Maximization and Gibbs sampling procedures as part of PhD), topic modelling
Interests	machine learning *esp.* in unsupervised learning using Bayesian modelling (not restricted to text analytics), Distributed semantics, Data mining, information retrieval, machine translation
Languages	English, Tamil (fluent), Kannada, Hindi

Experience

Sayint.ai (Zen3 Info Solutions)
Hyderabad, INDIA

SR. SOFTWARE ENGINEER (NLP) — *May. 2017 - till date*

- Contributing towards building a speech analytics platform called "Sayint.ai" for call-centers.
- Built NLP postprocessor for ASR outcomes using Thrax and NLTK.
- Built end-to-end analytics solution, which involved the functionalities of computing the speaking-rate of user, crutch-word rate, listen-to-talk ratio of Agent vs client in dual-channel scenario, identify longest monologue, detect speaker emotion from voice and text data, verify compliance w.r.t different call-center metrics – which involved building supervised/unsupervised classifiers. All of this involved goal-driven research activities.
- Help interns learn

Trinity College, Dublin

Dublin, IRELAND

DEMONSTRATOR DURING PHD TERM *Sep. 2013 - Sep. 2015*

- Demonstration involved interacting with students during lab sessions to help them with the problems they come-up in completing their assignments. The work also involved evaluating students assignments.
- Demonstrated for various modules such as Compiler design and computer programming.
- Tutored "Advanced computational linguistics" course for individuals which mostly involved unsupervised approaches to tasks such as Machine translation and topic modelling.

MagnaInfotech

Dubai, UNITED ARAB EMIRATES

ASSISTANT MANAGER (NLP) - CONTRACT ROLE *May. 2013 - Jul. 2013*

- Worked for **Emirates** on behalf of **GENPACT** with a contract from MagnaInfotech
- Worked on a *Proof of Concept* (PoC) to predict the spare parts that are likely to fail in the next flight cycle. The project involved challenges of extracting useful information from the manuals provided for the aircraft by the manufacturer.
- Then in identifying *aircraft on ground* (AoG) situation, there were multiple signals involved in governing that situation. Such signals in combination with the useful information extracted from manuals were used as features for a statistical machine learning system to predict first the situation causing AoG and then predict the spare parts causing AoG situation. This is more of an *outlier* detection task.

IB Technology (now Incedo Inc.)

Gurgaon, INDIA

TEAM LEADER (NLP) *Jul. 2012 - May. 2013*

- Worked for **Ditech Networks** (now part of **Nuance communications**) on the *PhoneTag* (Voice-to-Text) service.
- In the voice-to-text service, the work involved converting voice mail to text using a continuous speech recognizer (CSR) and the text obtained was further processed using a post-processor module involving NLP techniques to be converted to user-readable text.
- Used NLP rule-based techniques in formatting text such as identifying the right text to be capitalized, identifying the right place to punctuate, identifying the words to be converted to alphanumerics such as phone numbers and door number. Then identifying words relating to email, website and converting them to the right formats.
- Further, I was involved in rule-based named entity recognition. The different named entities that were identified include Person, Time, Date, DateTime, Money, Email, Website, Address and Place. All the work is in three different languages English, Spanish and French.
- Although, this rule-based system was in place serving customers initial work was done in moving this rule-based system to a statistical one, by generating training data for different functionalities.
- Responsibilities involved research and development of new functionalities to provide intelligence to the system and improve the output, bug fixing and coming up with new implementable ideas to make the current system work better with some team management.

Talking Heads Language Service

Sheffield, UK

INTERPRETER (PART TIME) *Oct. 2011 - Feb. 2012*

- Involved in interpreting and translation from English to Tamil and vice versa for the clients of Talking Heads which majorly included Solicitors, Doctors and Teachers.
- On ad-hoc basis I still translate documents.

WIDEX International

Sheffield, UK

STUDENT RESEARCH (WAS PART OF MSC MODULE FOR CREDITS) *Jan. 2011 - Apr. 2011*

- Research conducted to improve the speech intelligibility in the hearing aids for hearing impaired, titled 'Conclear: Enhancement of Speech Intelligibility', which was proposed by a Danish researcher from 'Widex'. The project was implemented in MATLAB, which involved customizing a GMM (Gaussian Mixture Model) classifier to identify the vowels and the lexical stress from the speech. On identifying the vowels and the lexical stress from the speech, the vowel regions and the stress regions were modified to get two different outcomes.
- I was particularly involved in the identification and modification of lexical stress. The identification of the lexical stress regions from the human speech involved training the GMM classifier using the annotated conversational speech data from TIMIT corpus. The modification of the speech involved exaggerating the identified stress regions based on a distance metric. This role also involved testing the outcomes, which involved subjective testing and objective testing.

Evalueserve

Gurgaon, INDIA

RESEARCH ASSOCIATE (PATENT ANALYTICS) *Jan. 2008 - June. 2010*

- Handled various projects involving patent drafting, drafting defensive publications, patent-ability / novelty searches, patent invalidation / validity Searches, enforcement/ EoU/ claim chart analysis in the domains of computer science, electronics, wireless technologies and encoding technologies.
- Also handled projects on competitive intelligence (patent landscape and patent portfolio) studies which required in-depth analysis of patent portfolio of the client's competitors.

Publications

Dynamic Generative model for Diachronic Sense Change Detection Martin Emms, Arun Jayapal published in *Proceedings of COLING 2016*, *Dec. 2016* — A Gibbs sampling algorithm is developed for a diachronic model to estimate the parameters from raw time-stamped n-gram data with no sense annotation. We are able to demonstrate the inference of parameters which plausibly reflect the objective dynamics of sense use frequencies, in particular the emergence of a new sense.

An unsupervised EM method to infer time variation in sense probabilities Martin Emms, Arun Jayapal published in *ICON 2015: 12th International Conference on Natural Language Processing*, *Dec. 2015* — We have developed an EM algorithm to estimate the parameters from raw time-stamped n-gram data with no sense annotation. We are able to demonstrate the inference of parameters which plausibly reflect the objective dynamics of sense use frequencies, in particular the emergence of a new sense.

Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners Erwan Moreau, Arun Jayapal, Gerard Lynch, Carl Vogel published in *PAN proceedings on Author verification, CLEF 2015*, *Nov. 2015* — Provides our system description for the author verification conducted by CLEF workshop.

Detecting change and emergence for multiword expressions Martin Emms, Arun Jayapal published in *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), EACL*, *Apr. 2014* — concerns unsupervised means to detect an emerging new 'MWE' usage of a given n-gram, such as the usage of smashed it to mean doing something excellently.

TCDSCSS: Dimensionality Reduction to Evaluate Texts of Varying Lengths-an IR Approach Arun Jayapal, Martin Emms, John D. Kelleher published in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), COLING*, *Aug. 2014* — provides system description of the cross-level semantic similarity task for the SEMEVAL-2014 workshop. I have used a dimensionality reduction technique for this task and this work was guided by Martin and John.

Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm Erwan Moreau, Arun Jayapal, Carl Vogel published in *PAN proceedings on Author verification, CLEF 2015*, *Sep. 2014* — concerns system description of Author verification task conducted by PAN for CLEF.

Vector space model and Overlap metric for Author Identification Arun Jayapal, Binayak Goswami published in *PAN proceedings on Author verification, CLEF 2013*, *Sep. 2013* — concerns system description of Author verification task conducted by PAN for CLEF. The system was developed using the very simple vector space model and worked reasonably well (just better than the baseline).

Similarity Overlap Metric and Greedy String Tiling for Plagiarism Detection Arun Jayapal published in *PAN proceedings on Plagiarism detection, CLEF 2012*, *Sep. 2012* — concerns system description of Plagiarism detection task conducted by PAN for CLEF. The system was judged first in the document retrieval task (one of two tasks involved in Plagiarism detection system).

USFD at KBP 2011: Entity linking, slot filling and temporal bounding Amev Burman, Arun Jayapal, Sathish Kannan, Madhu Kavilikatta, Ayman Alhelbawy, Leon Derczynski, Robert Gaizauskas published in *KBP 2011 proceedings on Knowledge base Population for Text Analysis Conference*, *Jan. 2011* — concerns system description for the Question answering task and this was part of my Master's dissertation. My role was to come up the Named entitity recognition (NER) model and co-reference resolution model. For NER, I used stanford CRF classifier and for co-reference resolution, I had come-up with a rule-based classifier.