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 | https://github.com/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 (Cmputational 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*

BE 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. For a detailed read, please checkout at `https://arunjeyapal.github.io/docs/thesis.pdf`

Skills

Programming	C++, JAVA, Python (2.7 & 3.5+), R, MATLAB, LaTeX
NLP/ML tools	NLTK, Thrax, scikit learn, pytorch, stanford coreNLP, etc.,
Databases & Tools	PostGRE Sql, Celery (microservice), Flask (build apis)
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

Honors & Awards

International

2013 **Awarded 66000 Euros grant at CNGL**, I was awarded 22,000 euros grant for 3 years for pursuing PhD at Trinity college Dublin and the grant was extended for another 6 months after the completion of 3 years. — *Dublin, Ireland*

Domestic

2020	**Received Pursuing Innovation Award at EY**, I was awarded for coming with and implementing innovative ideas in solving the HTS code classification problem	*Bangalore, India*
2020	**Received Extra Miler Award at EY**, I was awarded for identifying critical bugs in the development of a document intelligence model	*Bangalore, India*

Experience

EY (Ernst & Young) GDS

Bangalore, INDIA

Supervising Associate — *Nov. 2019 - till date*

- Responsible for building AI Quality assurance framework from the QA principles (1) Robustness (2) Fairness (3) Explainability. Additionally, developed best practices during the developmental/experimental stages of AI models
- Leveraged existing deep-learning models, to build classifiers
- Built end-to-end NLP system to assign HTS (Harmonized Trade System) code for any given product description. This was done by leveraging multiple sources of text data

Sayint.ai (Zen3 Info Solutions)

Hyderabad, INDIA

Sr. Software Engineer — *May. 2017 - Oct. 2019*

- Contributing towards building a speech analytics platform called "Sayint.ai" for call-centers. Developed and maintained all the below mentioned functionalities/platforms.
- Built NLP postprocessor for ASR outcomes using Thrax and NLTK. This involves denormalizing the text outcomes into human readable form eg., email uttered as arun dot jayapal at gmail dot com would be transformed to arun.jayapal@gmail.com
- 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. All the above mentioned functionalities were developed to score a contact center agent.
- Came up with algorithm to verify the agent script adherence; this involves verifying whether the contact center agent adheres to a/list of scripts provided. This further involved a scoring mechanism to score the agent at call.
- Developed a rule-based call sentiment analyser platform to provide positive and negative scores for sales calls in the travel domain, however this was built as a platform to be re-utilized for all other domains.
- Came up with supervised classifiers to identify sales vs customer-service vs enquiry calls; this was done by first categorizing the call segments (multiple classifiers were built) under different buckets and came-up with rules utilizing the segment-level classification to further classify the calls.
- Built an end-to-end email-bot solution to auto-respond email queries (from employees who does blue-collar job) directed to HR in an organization. This work required building upto 50 classifiers to answer different intents (the email queries were free-text and did not adhere to any particular format). This also required extracting named entities from the text – this was needed only for certain intents.
- Developed a Named entity recognition (NER) engine which involved annotating, training and testing models using CRF classifier and adapting BERT models for the same task. Additionally, this was made production ready by building API integrating with Celery framework. The NER was accomplished to identify personal information of the clients spoken on the call such as Name, age, gender, credit card number, post code, etc., and mask them so as to maintain privacy of the customers and comply with the security standards of different countries

Trinity College, Dublin

Dublin, IRELAND

Demonstrator during PhD term — *Nov. 2019 - till date*

- 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 were 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

For a detailed list of publications, please visit `https://arunjeyapal.github.io`