UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

COM61004: Introduction to Collaborative Research Practice for SLT ePortfolio

Student Name	Aaron Fletcher
Student Registration	230116573
Cohort Number	5
Academic Year	2023/24

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Contents

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Mini-Project

Each student will be expected to include
- *an individual log of contributions/activity organised on a chronological basis, i.e. a brief weekly summary of activity and their individual contributions. See the information about* worklogs *for additional guidance.*

Note: the group report will be submitted separately.
Add additional tables as required.

Your group	Team 1

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Work log - week 1

Date (w/b)	23/10/2023 (2 ½ hours)

Narrative	- 23/10/2023 (1 hour)
	  - An initial meeting to discuss data generation was held. This was a freeform meeting where people brainstormed ideas for generating the data. Little consensus was made, however, we have scheduled a meeting for later in the week to address some questions raised - what menu items to include, different scenarios/scripts to produce audio recordings, and what participant characteristics we want to cover. A Trello board with tasks and meeting notes were created in a shared file.
	- 27/10/2023 Second meeting based on an agenda (1 & ½ hours).
	  - I have taken somewhat of a role of project chair/manager by facilitating the meetings to ensure that people's opinions are heard. Various items were discussed based on the agenda, and tasks were assigned to subteams to meet again to discuss findings. I have been tasked with creating the agenda for the third meeting and proposing a data management plan with Boxshaun and Yao - this is an area that I have little experience in (besides using GitHub for my projects), so I will investigate the university's advice on this and discuss with my colleagues.

Work log - week 2

Date (w/b)	30/10/2023 (4 ½ hours)

Narrative	- 31/10/2023 Sub team Meeting (1 hour)
	  - First meeting with sub-team, where an overall approach to the topic was outlined, with key components such as ASR, Named Entity recognition separated into two sub-groups. Each sub-group will work on researching SOTA approaches for their small task before passing it on to the other sub-team to verify.
	  - *Task*: Research Named Entity Recognition and provide SOTA options for the project.
	- Data Management Plan (1 hour):
	  - In conjunction with the other DMP team members, a data management plan was outlined and will be sent to the library for review before submission.
	  - *Task:* Submit data management plan for review
	- 2/11/2023 Dataset recording Meeting (1 hour)
	  - Introduction by Stefen to the equipment that is available for us to use.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

	- 3/11/2023 Data Generation Team Meeting (1 & ½ hours)
	    - The project has changed the scope to focus more on data generation than previously, as approximately 80 hours of recording will be needed for publication.
	    - The preliminary cut-off point for 14th February is set for finishing data generation.
	    - Discussion of continuing ethics application.
	    - Preliminary discussion of how the data will be generated/Experimental design, i.e. closed microphone on the server + customer, with an outside array of microphones to provide a differential.

Work log - week 3

Date (w/b)	6/11/2023 (3 hours)

Narrative	- 9/11/2023 - Sub-team meeting
	- The Data Management Plan was submitted for review by the library team / Stuart Wrigley , and feedback was received. A few changes were made based on this - such as not using Google Drive to mirror the backup, as the university research storage offers more storage and is backed up on multiple sites. This is a live document so further iterations will be made depending on requirements.
	- Following a meeting with the CDT management team, the project scope was scaled back - from 80 to 10 hours of training data.
	- The data management plan was finalised by all DMP team members on 9/11/2023. This will be submitted on Monday 13/11/2023.
	- Me and FP created a Gantt chart for the Project management meeting, which outlined key project phases and their ending times.
	- Subteam work:
	    - A further 1-hour meeting was held where we discussed our research. I had been focusing on a potential named entity recognition. I created a list of potentially interesting NER models to investigate based on the f1 score and the accuracy of hugging face models. Collectively, we were interested in using span markers (a NER model which, rather than categorising the word, does so over a range of words related to the category, which might be more applicable to our applications).
	    - Decided to attempt to implement one of these models with different encoders, as it will form the research question of what encoder (tokenizer) results in better performance with span NER tagging.
	    - To enable comparison of our results, we will attempt to use the same dataset (Berkley), which we will annotate with our additional class.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Work log - week 4

Date (w/b)	13/11/2023 (3 hours)
Narrative	Sub-team meeting: ● Discuss what we plan to present in the mini-project meeting catch-up, as the team has submitted into three parts (NER, ASR and Query Detection); each group will contribute a slide on their work so far. ● I have been attempting to generate some training data, and some testing data which represent the problem more and code up a way of splitting these up. Data generation meeting: ● The meeting was interrupted by a fire alarm! ● Reduced scope to address issues with time in producing enough data (40 hours down to 10 hours).

Work log - week 5

Date (w/b)	20/11/2023 (4 hours)

Narrative	Sub-team (*4 hours*):
	● Formation of a slide deck for presentation for a mini project meeting.
	  ○ Included proposed task topology, work undertaken by the ASR subteam, the NER subteam, and the query identification subteam.
	  ○ Included metrics of current approaches so far.
	● I uploaded a working example of a Packed Levitated Marker for Named entity recognition to the group Github, which got 74% accuracy. This was trained on the FoodBased NER corpus and tested on the Berkley Restaurant Corpus.
	● Production of Gantt Chart for subteam.
	
	Data generation meeting - No meeting was held this week; however, some discussions were held in the Mini project meeting (*30 minutes*):
	● Further reduction of data gathering scope.
	● Discussion of experimental design.
	● Discussion of ethics application.
	
	Training in the Kroto lab (*30 minutes*): Health and safety induction into the Kroto laboratory now enables us to get some quick data to assess.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Work log - week 6

Date (w/b)	27/11/2023 (3 hours)

Narrative	Data generation progress (1 ½ hours):
	● Meeting 5 (*1 hour*)
	    ○ We discussed publication, with the consensus being that those wanting to publish will continue following the cessation of the mini-project.
	    ○ Established that we would want to record in the Kroto lab with scenarios. Dummy data is pending and will be recorded and shown at the next meeting.
	● Speech transcription and annotation meeting (30 minutes):
	    ○ Meeting with Minghui, established some brief outlines of what we want the annotation guidelines to look like (i.e. span tagging, item categorisation).
	    ○ AF to investigate program platforms to use this.
	Subteam progress (*1 ½ hours*):
	● Finalisation of subsections and upload of completed code to GitHub
	● Relocation to subset task to intent classification.
	    ○ Established the types of classes that we want to detect (Add, Remove, Edit)
	    ○ Established a dataset that we will use to create mock models:
	        ■ https://huggingface.co/datasets/xjlulu/ntu_adl_intent/viewer/default/train?q=food&p=1
	        ■ https://huggingface.co/datasets/silicone/viewer/maptask/train
	    ○ Once completed, I will attempt an LSTM approach to the data and report findings to the subteam.

Work log - week 8

Date (w/b)	4/12/2023 (3 hours)

Narrative	● Data Generation Team:
	    ○ Meeting with whole group (1 hour).
	        ■ I presented annotation work done since the last meeting, which raised some issues with their creation.
	    ○ Meeting with Stefen (30 minutes).
	        ■ Discuss some easy changes to the experimental design that will improve the data quality, and arrange training sessions in the Kroto Lab.
	    ○ Commence creation of annotation guidelines (1 Hour)
	        ■ Discussed aspects of annotation guidelines and ultimately decided that span marking was better for the data type we

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

	would be using. We are planning on annotating on three levels
	   ● Food type
	   ● Query Type
	   ● If to include within the dataset.
	■ Different platforms were discussed to do this (Prodigy vs Borcorro). Ultimately, we decided to use Prodigy, AF + MZ will create annotation guidelines for the rest of the group and present it.
	● Sub teamwork (30 minutes)
	   ○ Implementation of the intent classification LSTM model.

Work log - week 9

Date (w/b)	11/12/2023 (5 hours)

Narrative	● Data Generation progress (*4 hours*)
	   ○ Pilot recording in the Kroto Labs, including improving experimental design (2 hours)
	     ■ Meeting with Stefan Goetze and other subteam members to set up an experimental design within the Kroto laboratory that closely matches real-world conditions.
	   ○ Prepare a slide deck for the data generation and sub-team meeting with project directors, and attend the meeting (2 Hours).
	     ■ Discuss the current project status.
	     ■ Discuss potential changes to the annotation guidelines.
	     ■ Share pilot recordings and compare the two experimental designs.
	● Subteam Progress (*1 hour*)
	   ○ Creation of LSTM approach to intent classification for open-sourced dataset (1 Hour).
	     ■ 150 class classifiers which performed poorly (accuracy ~10%)
	     ■ 12 class classifiers which performed well (accuracy 100%)
	     ■ Updated to the shared GitHub repository.

Work log - week 10

Date (w/b)	18/12/2023 (6 hours)
Narrative	● Data Generation progress (*3 hours*): ○ Further pilot recording in the Kroto Labs, including improving

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

experimental design (3 hours)
- Generate pilot data, which can be used to further subteam approaches.
● Subteam Progress (*3 hours*):
  ○ A sub-team meeting (20/12/2023) will be held to discuss individual tasks over the holiday period and the formation of the research questions. I will focus on prompt engineering (1 hour)
  ○ Created a Python program to programmatically manipulate the recorded files and run various models on them, such as whisper, including prompt engineering (2 hours).

Work log - week 11

Date (w/b)	25/12/2023 (0 hours)
Narrative	● Holiday week

Work log - week 12

Date (w/b)	1/1/2023 (5 hours)

Narrative	● Subteam Progress (*4 hours*):
	  ○ Research question formation (3 hours):
	    ■ Formation of the prompt engineering research questions
	    ■ Generated 5-6 research questions on prompt engineering, including background research, assessing implementation and write up.
	    ■ RQ1 - Does providing a neogeolistic dictionary of words used in the transcription reduce WER for transcription.
	    ■ RQ1 - Does providing longer term dependencies reduce WER for transcriptions
	    ■ RQ3 - Does providing context improve accuracy?
	    ■ RQ4 - Does providing more context improve accuracy?
	    ■ RQ5 - Does chain-of-through prompting improve accuracy?
	  ○ Subteam meeting (1 hour)
	    ■ Meeting to discuss research questions, and explain to the group and discuss any potential challenges.
	● Data Generation progress (*3 hours*):
	  ○ Creation of annotation guidelines (1 hour)
	    ■ Started drafting more formalised annotation guidelines within a shared google document.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Work log - week 13

Date (w/b)	8/1/2023 (3 hours)
Narrative	- Subteam Progress (*3 hours*) - Creation of Research questions presentation - The slide deck was created in the shared Google Drive in anticipation of the meeting on 16/1 outlining the project proposal (*1 hour*). - Meeting with sub-team to discuss project proposal in person on Friday (*1 hour*). - Further refinement of project proposal document (*1 hour*)

Work log - week 14

Date (w/b)	15/1/2023 (3 hours)
Narrative	- Subteam Progress (*2 & ½ hours*) - Directors Meeting 3 (*1 & ½ hours*) - Preparatory meeting before presentation to revise approach (30 minutes). - Meeting with directors to outline project scope and email scoping document to all attendees (*1 hour*). - Investigation and discussion of finalised order evaluation metric (such as graph edit distance, etc) (*30 minutes*). - Generate some proof of concepts within Python to discuss with the group. - Data generation progress (1 hour). - Advise the rest of the group to review/edit annotation guidelines - Decide on purchase items for research. - Update re: ethics application.

Work log - week 15

Date (w/b)	22/1/2023 (3 hours)
Narrative	- Data generation progress (4 hours). - Finalisation of annotation guidelines (1 hour) - Recording within the Kroto laboratory for more pilot recordings (2 ½ hours) - This pilot recording identified some shortcomings with the research design, such as: - The Research Coordinator needs to be more active in organising the research and ensuring quality recordings.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

<table>
<tr><td></td><td>

An outline needs to be used for the server flow to ensure that key points are being hit.
The menu ordering system needs to be updated to current guidelines.

Generation of a research working log to keep participant characteristics (½ hour) and assign unique UID to each scenario/participant. Updated this to ensure that it is easy to use.

</td></tr>
</table>

Work log - week 17

Date (w/b)	29/1/2023 (5 hours)
Narrative	● Recording audio scenarios (5 hours)

Work log - week 18

Date (w/b)	5/2/2023 (6 hours)
Narrative	● Recording audio scenarios (5 hours) ● Preparation of Director meeting notes (1 Hour)

Work log - week 19

Date (w/b)	12/2/2023 (7 hours)
Narrative	- Annotation masterclass meeting (1 hour) - A meeting presented by Minghui outlined the generated transcription and annotation guidelines and how to use the program Prodigy to transcribe the audio. - Initial transcription test run (1 hour) - Tests were run where each person annotated transcripts to compare experiences and common challenges. - Post-run meeting (1 hour) - The decision taken at this meeting is not to annotate food/quantity, and the intention to speed up this process. - Clarification of common issues, such as overlapping speech, which we decided not to translate. - We also decided not to transcribe dysfluencies. - Subteam Meeting (1 Hour) - At the subteam meeting, we discussed scope reduction (removing the granular approach). - The code was generated for the prompt engineering approach. - Creation of the prompt engineering approach code (2 Hours) - Creation of the code to run prompts through gemini + gpt, with test runs being created via the whisper transcription directly. - To determine evaluation metrics.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

	● Transcription time (1 hour) ○ Transcription of the audio following above guidelines for this week

Work log - week 20

Date (w/b)	18/2/2023 (*4 hours*)

Narrative	● Creation of the prompt engineering approach code (*2 Hours*)
	  ○ Creation of the code to run prompts through gemini + openai, with test runs being created via the whisper transcription directly.
	  ○ To determine evaluation metrics.
	● Studio recording (*2 hours*)
	  ○ Recording more scenarios.

Work log - week 21

Date (w/b)	25/2/2023 (*3 hours*)

Narrative	● Transcription (*1 Hour*)
	  ○ Production of 10transcripts using prodigy
	● Prompt Engineering Coding (*2 Hours*)
	  ○ Further coding to complete this area

Work log - week 22

Date (w/b)	18/2/2023 (*3 hours*)

Narrative	● Project Meeting (*1 Hour*)
	  ○ Project Meeting with Stuart and Rob Gaizauskas to update them with the status of the project. Stark reminder that there are a few number of weeks left to produce the output (before easter!)
	● Meeting with Prompt Engineering Subteam (1 hour)
	  ○ Meeting with PE subteam to ensure that we are all on the same page with expectations set and deliverables mentioned.
	● Transcription (*1 Hour*)
	  ○ Production of 7 transcripts using prodigy

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Work log - week 23-24

Date (w/b)	18/2/2023 (*12 hours*)
Narrative	- Run Experiments (*4 Hour*) - Run experiments for the PE team using code created throughout this process. The experiments are listed in the final project report. - Meeting with Prompt Engineering Subteam (30 Minutes) - Meeting with PR subteam to clarify what the prompts are going to be (Me and FP) - Project write-up (*7 hours 30 minutes*) - Collaborative write up of the project using Overleaf/Latex with the team.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club

Each student will be expected to record in their e-portfolio:
- *their commentary bullets about each paper*
- *brief write-up of outcomes of the discussion from each session They should indicate if they were the discussion chair*
- *A copy of the slides used at each session for which they were the presenter.*

Add additional tables as required.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 1

Paper title	Speech Processing for Digital Home Assistants: Combining Signal Processing With Deep-Learning Techniques (2019)
Presenter	Jason Chan
Date	16/10/2023
Commentary bullets	**OVERVIEW:** Excellent starting point for students exploring speech concepts. **TARGET AUDIENCE:** It is vague; the paper presented a (kind of) systematic review and challenges within speech processing, yet it is in a magazine-style format. Readers are likely already engaged in speech processing, so is an "entry point" for this audience somewhat redundant? **SOTA:** Produced in 2019, so likely not. **ASSUMPTIONS:** ● Communication of information is only TTS; this is increasingly not the case as digital home devices increasingly have visual components, e.g. Google Hub. **WOULD HAVE LIKED:** ● An explicit discussion of the effect of multiple ASR systems in one environment. ● Discussion on how communication with these devices could be more natural and prescriptive. This could stem from the vagueness of what A "digital assistant" is. I.e. using "wake words" does not seem natural and changes my interaction with the device. **KEY POINTS:** ● Deep neural networks are not a universal solution for speech-processing tasks. ● Speech processing for digital assistants has yet to be solved. ● It is a great reference for speech processing challenges and (some) solutions.
Narrative	In this journal club, I was chairing the discussion. The presenter, Jason, covered the main aspects of the paper, outlining all the approaches given to speech recognition within the paper (Acoustic environment, Multichannel speech enhancement, automatic speech recognition, Text-to-speech synthesis and fully handfree interaction. This journal club reviewed a mini-review paper on approaches to key speech processing challenges circa 2019. Multiple points were raised during the discussion, such as who this paper was aimed at (it was concluded signal processing specialists who were interested in deep learning approaches). Some audience members opined this made the content less relatable to themselves, as the paper heavily relied on mathematical notation specific to the domain. Additionally, the magazine format made figures and references appear distant from the relevant text, hampering their understanding. While I did not necessarily agree (considering it was meant for production in a magazine, where it would have felt more appropriate), greater consideration by the authors of the main method of article consumption could have been given. **Action points to improve chairing:**

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

	● During the discussion, some people were more vocal about contributing than others, which, as the chair, is my role to adjudicate. To improve this, I asked targeted questions to people who had yet to have the opportunity to speak so far, ensuring that every person had contributed to the discussion. ● Initially, I would have people read their points from the slide deck; however, this created a very prescriptive approach to the discussion (people were waiting their turn, etc). The discussion was better suited to more of a conversation, with my prompt questions allowing people to add to the discussion further. Going forward, a more informal approach is recommended. ● Overall, it was a positive experience, which I require more experience with.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 2

Paper title	An Overview of the SPHINX-II Speech Recognition System.
Presenter	Anthony Hughes
Date	31/10/2023
Commentary bullets	**Positives:** ● Liked the introduction of key terms (Senones, etc). ● Seems like a big ASR development at the time! ● Uses a hybrid approach, which aligns with my cognitive systems bias on general AI! **Critiques**: ● Multiple instances of abbreviations are used without instantiating first (e.g. MFCC, HMMs, LPC). ● When introducing a novel term (Senone), a worked example is appreciated! ● A clear overview of the SPHINX system is challenging due to self-citation and domain-specific terminology. ● The paper only addresses out-of-vocabulary utterances beyond increasing the amount of data. ● Language models used are of their time (i.e. better options exist now than n-grams!). ● No consideration of noisy environments given - Would expect senone's will change with noise! **Questions:** ● What exactly is a semi-continuous hidden Markov model? ● Have hidden Markov models just been replaced by DNNs? ● Are some senone's language unique? ● Where are the codebooks? Could the author have provided more information or made them open source? Is this due to sponsorship restrictions?
Narrative	● Paul Gering chaired this journal club meeting. ● This journal club had a slightly different approach than the previous one, with the facilitator (Anton) interjecting more than the previous session (to clarify issues with our understanding of the texts). ● A key question was the motivation behind reading this paper, which seemed to be divided - motivations raised from it being a requirement for the CDT, reading because it is a foundational system of the topic area, and some people hand little motivation. This is important because, depending on the motivation, you will have different expectations of the article. ● Was a conference paper, which might explain the formatting as typically, abstracts are submitted before the article is written. ● Commercial application of SPHINX-2 was a quite successful dictation system.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 3

Paper title	From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation.
Presenter	Yanyi Pu
Date	17/11/2023
Commentary bullets	- Very well laid out, structured referenced paper for word meaning representations. Provides a great starting point! - Not fully comprehensive, but as stated (due to the pace at which the field evolves), it is not intended to be. Would have liked search methodology, however! - Great explanation of how polysemy affects word embeddings and research done to overcome this. - Translational-based embeddings - This is a new concept; how would it deal with words without direct translation (i.e. Waldeinsamkeit)? - Four axes, not three, on pg 481. - Do humans discriminate between words linearly? "Curse of multilinguality" - any information on why this is happening? Do people experience this curse? - Reporting Bias - Paper gives a great way of formalising this concept. - Good section on current challenges, such as degeneration issues in contextual embedding spaces.
Narrative	- Jack Cox chaired the journal club. - It was a very interesting and informative journal club despite the journal covering a large body of work. - The paper reviewed the majority of work that had been done on foundational topics within word embedding distributional models, such as how to embed meanings into Euclidean space, sense-aware embeddings and semantic knowledge injection to word embeddings. The journal club discussed: - What exactly is meaning and understanding (and a discussion of formal semantic language understanding vs distributional language understanding). - A discussion of symbolic grounding and how language can be very specific to the person using their language and grounded in their understanding. - Limitations of distributional language understanding approaches include whether we linearly distinguish between words / semantic meanings, their biases on training data and their inability to generate novel concepts. - What an optimal approach would look like - semantic knowledge injection through combining the two approaches. - Translational-embedding models: this was a new concept to most, and generally, people thought it would be an excellent way to encode semantics into models. We discussed how far apart the languages would need to be to encode semantic knowledge sufficiently (it seems like it would have to be not close enough but not too far apart. - People's uses for the paper, with the consensus being a reference paper for word embeddings used to look up studies performed in the area, but it not being an exhaustive representation of this topic in the fast-moving domain.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 4

Paper title	Inter-Coder Agreement for Computational Linguistics
Presenter	Aaron Fletcher
Date	8/12/2023
Commentary bullets	- Full link here. - Does Inter-Coder Reliability Matter? - Low inter-coder reliability means unreliable datasets. - If two coders produce similar results, then they have a similar internal understanding of the annotation scheme. - Reliability is required for validity. - If coders are inconsistent, then - Annotation doesn't capture the truth. - Some coders are incorrect. - Paper Contribution: Common notion - Solution: Provide more! - items { i | i ∈ I } of cardinality i - categories { k | k ∈ K } of cardinality k - coders { c | c ∈ C } of cardinality c - Ao: Observed Agreement - Do: Observed Disagreement - Ae: Expected Agreement - De: Expected Disagreement $p(\cdot)$: Probability of a variable $\hat{p}(\cdot)$: Estimate of the probability from observed data - n with subscript to indicate the number of judgements of a type. - How to measure Inter-Coder Reliability - Agreement without chance correction. - Chance correlated coefficients. - Coefficient S κ - π κw - Krippendorff's α - Fleiss's Multi π - Multi-κ - BUT - Which measure do we choose and why? - K, what is it good for ? - Absolutely nothing :) - K is any kappa-like coefficient (κ, α, π etc). - No universally accepted standard. - Medicine: >0.4 - CL: > 0.8 - The authors: No specific threshold

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

- - - Krippendorf themself thinks > 0.9!
    - In ML/AI we are often generalising patterns, not hypothesis testing, and K is a poor predictor of ML success.
    - K scores are ultimately tests for soundness of annotation scheme.
- Bias and Prevalence
  - Bias problem: Π/κ different when annotator marginal distribution are wildly divergent.
  - Prevalence problem: When a most of the items fall under one category, might lead to artificially high observed agreement by chance.
  - Authors suggest coefficient choice should be based on desired interpretation of change agreement not magnitude of divergence.
  - In reality we typically have a small subset of data where c > 1, which reflects the reliability of the annotation procedure, so Π / α more appropriate.
  - κ provides more information information regarding validity - it rewards biased annotators.
  - Can express overall annotation bias as the difference between κ and Π.
- Unitizing
  - The process of identifying units of annotation (typically linguistic units e.g. words, utterances, or noun phrases).
  - Segmentation/Topic Marking: Portions of text that constitute a unit because they are about the same topic (Think of our mini-project, where we are marking "2 burgers" as one topic!).
  - Agreement coefficients with segmentation likely to be lower:
  - Boundary/not boundary distinction: K = 0.647
  - TREC segmentation of broadcast news: K = 0.784 / 0.36
  - Identification of argumentative zones: K = 0.81
  - Conversational games: K = 0.59
  - People tend to agree on the bulk of segments, but not the boundaries!
- Anaphora
  - Where a grammatical substitute relates to a previous word or topic.
  - Lots of different types (some authors suggest 12 types!).
  - E.g.
  - Each fall, penguins migrate to Fiji. It happens just before the eggs hatch.
  - it is referring to the migration of the previous sentence
  - M: ....
  - first thing I'd like you to do
  - is send engine E2 off with a boxcar to Corning to
  - pick up oranges
  - as soon as possible
  - S: okay
  - M: and while it's there it should pick up the tanker
  - Passonneau's Proposal (using sets of mentions of discourse entities as labels).
- Word Sense Tagging
  - Bank can refer to a financial institution or the side of a river.
  - "I deposited money in the bank", Bank should be tagged with sense of financial institution.
  - Requires dictionary: often coders will have different understandings of each word sense, and gets even more complicated with polysemous verbs.
  - Paper suggested use professional lexicographers and arbitration.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

| | ○ Use coarser grained classification schemes, which group together dictionary senses (Wordnet) ○ Call has 28 fine-grained senses in Wordnet 1.7 ● The Main Takeaways ○ Report your inter-coder agreement, the maths isn't that scary! ○ You can classify intercoder correlation in terms of agreement or disagreement. ○ Justify statistical test(s) beyond it is standard for inter-coder agreement! This paper is an excellent reference. ○ But if we were to generalise … ■ $c = 1 \rightarrow$ No need ■ $c = 2 \rightarrow \kappa$ ■ $c > 2 \rightarrow$ Krippendorff's α ○ Don't get hung up on "perfect scores". ○ More coders, better results. ○ How often do we have $c > 1$ in the age of "big data"? ○ Why did they not provide a worked example for when $c > 2$? ○ What is the "sweet spot" for variability? Often, we use noise to increase the robustness of our models. |
| **Narrative** | ● All participants well-received the paper, highlighting its usefulness as a reference when undertaking dataset annotation. ● The discussion focused on the variability of an acceptable K value and how it should be used. The consensus was that the actual value doesn't particularly matter. ● The group thought the conclusion could have been more explicit in their recommendations. ● The importance of expert coders was discussed. ● Is high agreement a requirement / relevant for the real world? ● [The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing](#) |

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 5

Paper title	Efficient Estimation of Word Representations in Vector Space
Presenter	Paul Gering
Date	5/2/2024
Commentary bullets	- Main focus on the paper is clearly explaining how these models are computationally efficient, and why previous models, such as n-gram modelling are not. - Further optimisation given through the use of Huffman trees / hierarchical softmax given with a clear explanation. - Doesn't delve into why CBOW scores higher for syntactic accuracy (64%) vs Skip-Gram on Semantic accuracy (55%). - Questions: - With CBOW why are we sampling future words? When writing I don't tend to consider future words within a sentence, rather I attempt to convey a thought. - Both skip-gram and CBOW use context windows, what are the optimal lengths for these windows? - Would it account for homonyms/polysemes - don't think so! - Why did they choose 640 dimensions? - Is King - Man + Woman = Queen a valid evaluation metric?
Narrative	This was an hour-long journal club which was chaired by Minghui Zhao. This paper covers estimation of word representations in vector space. - Seminal Paper was generally well received. - Very good replicability within the paper. - Does this represent learning and understanding, human need a few examples, and we generalise it very well whereas the neural networks do not do this. - Discussed assumptions of the models - distributional hypothesis, linear relationships, context window. Doesn't consider long-term dependencies. - Listed the vectors which were released. - Bias within the models, discussed if they represent understanding or understanding. - If you evaluate the distant/top k predicted things then you are introducing more assumptions and more hyperparameters. Which improves the clarity of the causes of this. - Circular analysis potential present, which could present train data contamination. - Would have liked more analysis of the results in the paper. - This was about the time that GPUs were starting to be created that could make use of neural networks. - Focus in NLP was very different in 2013, SVM, etc. - Table 4 - Uncertainty over whether the comparisons are valid. Issue if it was a valid comparison. This is a proof of concept rather than a benchmarking paper. How much of the information is useful in the table. - Document/ word level meaning representations were popular at the time, so vector representations were used as it was the goal, rather than natural language generation.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 6

Paper title	Efficient Methods for Natural Language Processing: A Survey
Presenter	Boxuan Shaun
Date	20/2/2024
Commentary bullets	- Reads like a cookbook on how to approach optimisation for NLP :). - Due to the attempt to cover everything, more is needed; however, the function of this paper is to provide readers with references on the various approaches to improving efficiency. - Would have enjoyed more discussion on pruning, particularly given how closely it closely matches what happens with humans - see synaptic pruning. - This very effective paper highlights that universities need more hardware capability to analyse these models. - No methodology was presented for their search strategy for papers (was this a systematic approach or cherry-picking of papers?) - Some newer approaches (prompting, knowledge distillation) are new and likely to be updated rapidly due to the field. Could it have been better to go more in-depth on these areas or potentially a series of papers? - Great general explanations of each optimisation concept! - I liked some practical examples of implementing some of these optimisation concepts, which might be out of the scope of this paper but would have made it a great resource! - Very little consideration for the interpretability of the models - does this have to come at the expense of efficiency?
Narrative	This was an hour-long journal club which Ian Kennedy chaired. The paper covers increasing the efficiency of the LLMs in NLP. - Participants discussed the breadth and context of this paper. It was noted that the paper came from a discussion in Germany. This potentially biases the reported papers to the participants of that discussion. - Due to the large breadth, there was limited discussion of each of the topics brought up in the paper. Some sections need to be more developed (such as the prompt engineering section), potentially owing to it being an emergent field at publication. - It provided a starting point for evaluating efficiency, yet a clear approach was needed. - Very little consideration of the increased interpretability of the models; does this have to come at the expense of efficiency? - Some participants didn't like the fact that opinion was not offered in the paper, while others offered too much opinion through the paper. - A significant issue with the paper was the need for a clear methodology for the search criteria. Was this a systematic review, or was this an opinion piece? What were the inclusion/exclusion criteria? - Discussion of the role of surveys and key differences between systematic reviews. surveys flirt too close to being systematic reviews to not explicitly state they have not effectively searched the search space, which can lead to misunderstanding.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 7

Paper title	Position information in transformers: An overview
Presenter	Minghui Zhao
Date	7/3/2024
Commentary bullets	- Great survey paper on the different approaches to position information in transformer models, it excelled at highlighting variations of a single aspect (position information). - Unified mathematical notation is great! - Table 1 and 2 are great contributions to area, as it is a valuable reference for future work. - Clear logical layout to the paper, particularly the recurring concepts section (Reference point, injection method, fundamentals). ¼ of a page of text explaining how a table is showing related papers (740)! Potentially could be more concise here. - Quite liked the point they made NOT to provide quantitative comparison for this survey, it would have been near impossible to recreate all these models on the same dataset, and evaluation metric. - Re: Character level processing - characters are semantically void, so would this have more of a downstream effect on polysemy tasks (assuming a non-infinite window length for positional information)
Narrative	This was an hour and a half long journal club which Tom Clark chaired. The paper covers the various approaches to position information within transformers. - Participants discussed the breadth and context of this paper. It was very focused on a small feature of transformers, which allowed it to be a great reference for information in that domain. - Discussion was had on why we need to encode positional information - for example, why order matters with text and noting that the transformer model is order invariant. - Participants discussed if positional information was even a pre-requestite for accurate understanding of sentences. Some argued that if we look at the start and the end of the sentences we can potential predict the entire sentence from that point (langauge modelling), and we do this as humans. I personally didn't agree with this, as we do not have proof on how humans encode information, and there is a potential infinite number of permutations that a sentence could take. I think positional information is important for comprehension by transformers. - Questions were varied from the group, with discussion on why they were investigating sinusoidal embeddings, despite them not having positional information. - Discussion on the different level of embeddings, such as character vs subword/word. I was interested in why character level embeddings might provide better results as they are semantically void, to which Anton suggested that it might be because of the fewer permutations that tokenizing at this level could bring, reduced out of vocabulary issues, and that they potentially could work better on smaller datasets, due to it producing more examples.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 7

Paper title	Position information in transformers: An overview
Presenter	Minghui Zhao
Date	7/3/2024
Commentary bullets	- Great survey paper on the different approaches to position information in transformer models, it excelled at highlighting variations of a single aspect (position information). - Unified mathematical notation is great! - Table 1 and 2 are great contributions to area, as it is a valuable reference for future work. - Clear logical layout to the paper, particularly the recurring concepts section (Reference point, injection method, fundamentals). ¼ of a page of text explaining how a table is showing related papers (740)! Potentially could be more concise here. - Quite liked the point they made NOT to provide quantitative comparison for this survey, it would have been near impossible to recreate all these models on the same dataset, and evaluation metric. - Re: Character level processing - characters are semantically void, so would this have more of a downstream effect on polysemy tasks (assuming a non-infinite window length for positional information)
Narrative	This was an hour and a half long journal club which Tom Clark chaired. The paper covers the various approaches to position information within transformers. - Participants discussed the breadth and context of this paper. It was very focused on a small feature of transformers, which allowed it to be a great reference for information in that domain. - Discussion was had on why we need to encode positional information - for example, why order matters with text and noting that the transformer model is order invariant. - Participants discussed if positional information was even a pre-requestite for accurate understanding of sentences. Some argued that if we look at the start and the end of the sentences we can potential predict the entire sentence from that point (langauge modelling), and we do this as humans. I personally didn't agree with this, as we do not have proof on how humans encode information, and there is a potential infinite number of permutations that a sentence could take. I think positional information is important for comprehension by transformers. - Questions were varied from the group, with discussion on why they were investigating sinusoidal embeddings, despite them not having positional information. - Discussion on the different level of embeddings, such as character vs subword/word. I was interested in why character level embeddings might provide better results as they are semantically void, to which Anton suggested that it might be because of the fewer permutations that tokenizing at this level could bring, reduced out of vocabulary issues, and that they potentially could work better on smaller datasets, due to it producing more examples.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 8

Paper title	An Overview of Speaker Identification: Accuracy and Robustness Issues.
Presenter	Jack Cox
Date	21/3/2024
Commentary bullets	- It is well structured and provides great justification for its need, with clear explanations of verification vs identification. - Why would you need to "balance" FAR = FRR, isn't this particularly application dependant - assuming you want higher FRR in say banking applications? - Features section is a good recap of information learnt in COM 6502 speech processing! - This tutorial covers a lot of ground, missing data, gaussian mixture models, universal background model, SVM etc - Might this had been better focusing on one aspect of these rather than every aspect. Do we really need basic principles of SVM separating hyperplanes for example (Figure 11 page 33)? - GMM-SVM/GMM-UBM perform better with mixture increase, so why does author state that "the importance of using a GMM-UBM approach when confronted with limited amounts of training data " - What am I missing here?
Narrative	This was an hour-and-a-half-long journal club that Anthony Hughes chaired. The main discussion stemmed from the purpose of this paper, in that it attempted to cover an extensive range of topics, yet crucially, it needed to provide more detail in each section. A good example is the necessity for discussion of support vector machines in this paper, which went into great detail despite not being the primary focus of this paper. People who were more interested in the speech technology side were interested in this paper, as it provides a good reference for fundamental topics (such as Fourier transformations etc. As someone uninterested in speech, this offered little exciting content for me. The paper's definition of speaker verification vs identification was not as clear for me as it was for other readers. After a brief discussion, we collectively concluded that this might be due to people not considering this a set problem (i.e., verifying against a known number of people for access vs verifying against potentially every person), hence why complexity would increase with the latter task. Stefan chose this paper for the journal club to expose us to a more foundational paper and to introduce us to approaches that we might not enjoy when writing papers. I took from this paper to not promise the premise of a tutorial when not giving exactly that, and also to ensure that my papers are more focused on a narrow area.

SLT

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 9

Paper title	Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond
Presenter	Fritz Peters
Date	15/4/2024
Commentary bullets	- Felt that the introduction was very underwhelming: - Failed to provide a clear concrete example of how causality is different from prediction (to the point where I had to look elsewhere!) - Assumed that causality is what would happen to Y if we change X, whereas prediction is knowing X predict the next Y. - Good description of confounding variables. "We cannot rely on the usual assumption that training and test data are identically distributed" - Besides the lack of clarity about what this means, does anyone actually assume this? - The "Blackbox metaphor" of DNN is tired and I am ready for it to never be published again - It actively paints DNNs as "unknowable" - Models are not trained for reasoning, just predictors. - When we make a cup of tea, we rarely reason why we put the milk in after the tea has brewed, we do it because we always have done. Only when it goes wrong, do we start to reason why it was done that way. Why is it a shock to anyone that models that predict do the same? . - Good mathematical background given, especially as someone who had little causality experience, along with clear mathematical notation. - Provides me with a new approach to looking at NLP and approaching it in a way that might lead to more interpretable results. - Style: Lots of words used to outline what will be explained (at some point…), rather than bait the reader, structure your content better.
Narrative	This was an hour long session with Fritz Peters presenting and Yao Xiao chairing the journal club. The presentation style was well given, with key points of the paper abstracted which really improved understanding. While this paper had a really interesting topic, the majority of participants did not enjoy how this paper presented arguments, with the consensus being interesting area, however poor execution. The authors did however present a convincing reason why casuality needs to be considered when creating NLP applications. A shortfall, I felt, was that saying that current NLP models don't consider this is a bit reductive as well, they are not designed to do this. As no comparative evaluation models were provided which showed the improvement that casualty has over correlational approaches, it still remains to be seen if this approach is better. Additionally attempting to shoe horn in causality as an improvement on correlational models wasn't fully evaluated. When we do things in real life we don't always consider causality, when say making a cup of tea, we rather rely on previous experiences to create predictions about the process. The paper would have benefited from more examples, particularly comparative examples between causality and correlation. It did however achieve its goal for us to consider modelling from a different approach.

Journal Club 10

Paper title	Self-Supervised Speech Representation Learning: A Review
Presenter	Yao Xiao
Date	13/5/2024
Commentary bullets	- A great resource for speech researchers interested in self-supervised speech representation Learning. - Had to use external resources to more clearly define the exact difference between semi-supervised and self-supervised learning. - Why use the term generative approaches when it already exists (and it widely used)? - How would contrastive learning approach multiclass problems? - Good demonstration of increasing dataset resulting in diminishing returns after a certain point/ choice of dataset etc. - Zero-resource speech technologies was interesting, as it is is obviously possible for speech to be learnt without a joint representation (i.e. people who are blind from birth), however in that scenario they still have other joint representations of the speech (just not text). - Great outline as to why speech problem is different to nlp.
Narrative	This was an hour and half long in person journal club presented by Yao, and chaired by Boxuan. Yao's presentation style was impressive, and she managed to reduce the more complex parts of the paper into an easier to understand format, without trivialising the content. The presentation itself provided me with an accurate definition of codebook vectors, which was something that had been previously brought up in the journal clubs but not fully explained - a fixed-sized table of embedding vectors learnt by generative models. What I found particularly interesting about the discussion was the contrastric predictive coding, which is a n -1 problem where you group your classes into whats included and excluded. This approach might not work with wave2vec which is a n binary classification problem. Furthermore determining true positive and negative samples with CPC would surely still be a semi-supervised machine learning approach, as it would need labeled instances to begin with. Yao had no questions on her presentation. In general the group liked the paper, and felt it provided a well structured. It was noted by Anton (the JC chair) that the maths had been very simplified in the paper as it is meant to provide a general approach to self-supervised learning, which at times, comes at the expense of accuracy. There was some people who did not like some of the tables, which were presented poorly. We discussed if multimodal joint representations is a prerequisite for unsupervised learning in speech (which, can be done by blind from birth people. We discussed the reasons for needing this type of learning, given that if you add in all the possible variations of the formants of speech with the general population, it quickly scales to enormous scale. We also discussed if layers represent specific learning, or if this is just us wanting there to be a pattern to this.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Journal Club 11

Paper title	Attention Is All You Need
Presenter	Ian W Kennedy
Date	13/5/2024

Commentary bullets	• For a given hyperparameter h, wouldn't the linear projection within the multihead attention of VKQ be very similar? Additionally, what is the effect of higher values of h (besides increased computational power)?
	• Are transformers just finite hidden state RNNs? If so, given unlimited computational power, would RNNs be more performant than transformers, as it captures more information about the hidden states?
	• This was a paper builds on what I would consider foundational papers for modern NLP, and was for people who are interested in NLP, and had a presumed background knowledge (i.e. Seq2Seq & Jointly Learning to align and translate, which was great for me!). Can see value for speech people in it to, considering lots of these approaches have been adopted by them.
	• Interesting comments by the reviewers in NIPs 2017.
	   ○ They mention the papers lack of statistical testing conducted on the ablation studies on hyperparameters.
	   ○ "While none of the underlying techniques here are strikingly novel in themselves, the combination of them and the details necessary for getting it to work as well as LSTMs is a major achievement".
	• While the title is meant to be catchy, is attention really all you need?
	• While it reduces the context vector bottleneck, it doesn't completely remove it, meaning that large context window documents would still be challenging (despite claims of infinite length context windows…).
	• What other features are needed to improve performance - reasoning, structural inductive biases?

Narrative	This was an hour-long in-person discussion with Jason Chan, who chaired the meeting. Discussion centred around whether the paper justified its choices well. While the paper was not intended to be a review paper, it did not have much analysis of the topics presented within the paper (such as why cos/sine was used over techniques such as CNN positional encodings or rotary positional encodings). While the paper conducted ablation studies on the number of heads used for attention, more discussion was needed about why this information performance decreases after a certain number of heads. Additionally, we discussed whether attention is the only mechanism used for the next token prediction, as it does not mimic well what humans do within sentence generation. The presentation style was short yet very informative, which was very suitable for the group that had previous knowledge of this paper, which is considered one of the most impactful papers in the field. The discussion also occurred on whether the issue of fixed context vectors being a bottleneck is fully solved with the multihead attention / positional attention mechanism.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Meta-MOOC

For each block, record
- *the topic you worked on (including links to the relevant Meta-MOOC platform pages) and the other person / people you worked with*
- *A brief description of your individual contributions to each topic you have been involved with, including:*
  - *a list of the resources (websites, videos, lecture slides, etc) you considered for inclusion in the Meta-Mooc on this topic plus an indication of which you chose to include and why, and which you chose to exclude and why*
  - *an indication of what additional material (annotations, comment, structuring) you contributed to the Meta-Mooc on this topic.*
- *Make sure you timestamp each part of the block narrative to show how your work built up over the 8-week block period*
- *See the information about worklogs for additional guidance*

Add additional tables as required.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Block 1

Topic	Speech & NLP in Healthcare
Group	Aaron, Fritz, Cliodhna, Paul
Dates	Before December
Slides	- Page: Challenges for Speech and NLP in healthcare. - Fixed broken links/grammatical issues. I renamed/restructured the page to address speech centricity. Five new challenges of NLP within healthcare were added. - Refactored existing linked programme, which produces up-to-date reviews of all publicly available clinical NLP.

Block 2

Topic	Dialogue Systems
Group	Aaron, Fritz, Jack
Dates	December to 23 February 2024
Slides	Completed Jobs: - Frame-based dialogue systems: merged with Dialogue-state architecture - updated/fixed inks - Added resources section - Dialogue System Issues - Created new page containing table with links to resources on ethical/implementation issues with dialogue systems - Examples of dialogue systems - Reviewed resources embedded within prose on Chatbots page - Created table incorporating these resources and my comments - Found and added resources relating to ChatGPT, and further resources on ELIZA Things to do: - Restructuring remaining pages to remove large blocks of text - Further resources in examples of dialogue systems, particularly for spoken dialogue systems - Expansion of dialogue system issues page

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Block 3

Topic	Prompting Methods in NLP
Group	Aaron, Ian, Tom
Dates	23rd February till 29th April 2024
Slides	Completed Jobs: ● Added new page under techniques outlining Zero/Few shot approaches ● Along with relevant FS/ZS papers, and video lectures found online. ● Under techniques, added links to the existing sub-pages, and links to four review papers on prompt engineering. ● Updated the work log for this page ● This page requires lots of work still, as the area is currently undergoing a lot of research! Things to do: ● Restructuring remaining pages to remove large blocks of text ● Reformatting Prompt Tuning and Calibration

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Technical Training

- *A list of all relevant training courses (e.g., HPC, RSE, programming, ML, etc) undertaken with evidence of completion. Include course title, provider, duration, dates.*
- *Reflection on lessons learned in the context of your group work and research project.*
- *A review of where the learned skills have been applied, with evidence.*

Add additional tables as required.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Training Course 1 - 2 ½ hours

Title	Basics of Supervised Machine Learning
Provider	Research Computing Training
Duration	2 ½ Hours
Dates	6/10/2023
Narrative	This training course was a lecture combined with practical elements, focusing on machine learning and supervised learning basics. It covered foundational aspects, such as features, target values, and classifying different machine learning approaches (supervised, unsupervised and reinforcement). The lecturer outlined how linear regression describes the data trend and how it can be used to predict quantities. They also presented classifiers as a way of distinguishing data points. The lecturer's high-level overview of how to approach machine learning tasks was appropriate and useful, and breaking each stage down and presenting the reasons behind each was informative. I found the amount of knowledge I had retained from my previous MSc module on machine learning and felt it helped me understand some of the more complex questions presented in the training course. The training course was aimed at people who may have yet to gain experience with the statistical methods presented however, as a recap, it felt sufficiently useful. The packages used were sklearn, which I had less experience with than tensor flow however, this was a relatively easy learning curve. Some topics I had previously learned were covered - such as ● Regression vs Classification. ● Data splitting for training/test datasets. ● Data normalisation and standardisation. ● Using trained linear regression models to predict new values. ● Use of evaluation statistics, such as R squared, MSE, etc. ● Concepts of overfitting/underfitting and its impacts on generalisation. New concepts gained from the session ● Importance of training R squared being similar to test R squared, and its value in generalisability. ● Multicollinearity is a concept I had a vague idea about; however, I can now understand how multicollinearity inflates values such as R squared, and using other metrics such as adjusted R squared would be better. ○ I.e. the more independent variables you add to a machine learning model, the more accurate your model becomes. The lecture/training style of the presenter was very good and inclusive. The session felt rushed at no point, and their use and interactivity were exceptional. Every person in the room contributed.

SLT

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Training Course 2 - 2 ½ hours

Title	Advanced Python: Language Features
Provider	LinkedIn Learning
Duration	2 ½ hours
Dates	15/10/2023
Narrative	This was an online lecture format, which was provided through the Linkedin Learning platform. I elected to do this training to improve my Python coding ability, which, while I think is currently satisfactory, I aim to improve the comprehensibility of my future code. The lecture format allowed for ad-hoc training, which I enjoyed, as I completed a section at a time, with a break in between. The lectures covered: ● Python coding style, current standards such as PEP 8. ● Some idiosyncracies of Python, such as strings vs bytes processing, boolean type casting and the walrus operator (an operator which can be used to increase comprehensibility of while loops). ● Advanced functions of Python, such as how to implement document strings (which, while I had previously understood, did explain its importance in helper functions) and lambda functions (a way of creating a small localised function which increases comprehensibility). ● Comprehensions (List, Dictionary and Set). I have previous knowledge of them but needed to understand their syntax fully. ● Classes of objects, such as enumerations (a way of binding values to a symbolic variable to increase comprehensibility and reduce computation), overriding unique methods, such as __repr__ and __string__, and how to enable object operations, such as obj1 + obj2 with overriding of comparison operations. ● Structural Pattern Matching, such as match, pattern guards, and sequence patterns. The lecture has taught me new Python features and will improve my ability to communicate with other researchers. I have started implementing features, such as list comprehensions in the lab sessions for each course I am on and conforming to a Python coding style.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Linked in Learning

Advanced Python: Language Features

Course completed by Aaron Fletcher

Oct 15, 2023 at 07:37PM UTC · 2 hours 20 minutes

Top skills covered

Python (Programming Language)

Head of Content Strategy, Learning

Certificate ID: 6db5fb616c2473e55bd597689c6c778645d071810b4d7fddeaf1d8674d055d6d

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Training Course 3 - 2 ½ hours

Title	Temporal Analysis in Python
Provider	Research Computing Training
Duration	2 ½ Hours
Dates	27/10/2023
Narrative	This was an in-person training session on the use of Python for temporal analysis of data. This is something that I was very keen to participate in, as my background is in medical clinical notes, which follow a temporal pattern; I was looking to gain further insight into this and how it might further my understanding in that area. The session started with a brief introduction to the problem, with examples being lifted from astrophysics. The majority of the terms outlined I had covered in my other module (Signal Processing), however, this formed as a refresher/revision for that. Some terms were not the same, however, that was because multiple key terms were different in each speciality. The session was highly interactive, with the presenter posing a mixture of theoretical and practical questions, such as the role of imputation/ how to deal with data gaps. From previous reading, this is quite a controversial topic, and the imputation of missing data (i.e. the act of determining a missing data value from related data) is an area that I need to research further. A key novel point addressed was that conversion into a stationary process (i.e. detrending through differencing) removes observations from the dataset, which can have unintended effects (such as removing a key factor in your analysis). The next part included a more intuitive understanding of the Fourier transformation, which is something that I will likely be using extensively over the next few months. I had previously explained this through mathematical terms, however, this presentation showed it through images and allowed me to understand the Fourier transformation graphs based on their input diagrams. A particularly impactful aspect of this training was that the presented actively asked questions, and if we raised questions, it showed us how to test what we thought through the use of Google Colab. Key topics covered: ● Temporal data, and how it differs from other types of data ● Stationary vs Non stationary data ● Fourier Analysis ● Hanning Windows ● Types of noise ● Practical applications of this through NLTK other packages within python.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Training Course 4 - 3 hours

Title	Supervised Machine Learning 2
Provider	Research Computing Training
Duration	3 Hours
Dates	15/1/2024
Narrative	This was an online training session provided by the research computing training at Sheffield University as a follow-up to the supervised machine learning I undertook in October 2023 (Basics of Supervised Machine Learning). The session covered the history behind the neural network, its attempts to approximate the human biology of the neuron, and early entries into the domain, such as perception. It covered basics such as the overarching topology of the domain and attempted to outline the differences between artificial intelligence and machine learning. The teaching catered to people with limited or no knowledge of neural networks and did not focus on the mathematical principles behind them. This meant the training was less relevant to me than formal machine learning training I had already undertaken. What was new to me was the TensorFlow playground, which the lecturer linked. It is an online toy program that allows users to add additional neurons to see its effects on a classification problem. It had different issues and succinctly outlined why solving non-linear classification problems with logistic regression is difficult. The lecturer gave problems to solve during the session, with work examples. The examples were relatively basic, yet they were a good introduction to using the tensorflow package, which I have limited experience with compared to the other packages, such as Pytorch. As other research papers can use Tensorflow, it was relevant to my overall PhD from this aspect. This was an online session, which felt less interactive compared to the in-person session and could discuss problems and ask questions. The lecturer answered the questions presented; however, due to the online nature, I got less answers to these questions and interactions. I still have questions regarding the impact of depth and width on neural networks and how to approximate the neuron by removing/ editing neurons not used to solve the issue (such as apoptosis from human biology). The refresher on activation functions felt relevant and useful. Key topics covered: ● Approach to classifiers vs prediction ● Hyperparameter optimisation ● Activation functions ● Supervised machine learning

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Training Course 5 - 2 ½ hours

Title	Python: Design Patterns
Provider	LinkedIn Learning
Duration	2 ½ hours
Dates	23/06/2024
Narrative	I recently completed a two-and-a-half-hour course on LinkedIn Learning that focused on design patterns in Python. I enrolled in this course to improve my ability to communicate my code effectively with others, especially because I am involved in open research. My code must be well-organized and easy to understand, ensuring accessibility for others. The training course covered various approaches to designing patterns within Python, such as creational patterns, structural patterns, and behavioural patterns. Creational patterns include factories, abstract factories, singleton, builders, and prototypes. Factories are used when you are uncertain about the types of objects the system needs or when your application needs to decide what is used at runtime. Aibstract factories take this further when users expect to receive related objects at runtime. Singletons are used when you only want one object to be created from a class and to share global variables, which allows multiple object instances to share the same state. Builder patterns attempt to prevent the anti-pattern of telescoping constructions (where an object has a complex number of constructors) and prototype clone objects according to a prototypical instance. Structural patterns include decorators, proxies, adapters, composites, and bridges. Decorators add additional functions to established functions or objects without modifying their code. Proxies attempt to postpone object creation in situations with high resource requirements and create objects when necessary. Adapters convert the interface of classes into another one the client expects. Composites maintain a tree data structure to represent part-whole relationships. Finally, bridges attempt to untangle unnecessarily complicated class hierarchies. Behavioural patterns include observers, visitors, bridges, and more. Observers can monitor for changes in the subject by constructing a one-to-many relationship between the subject and the observer. Visitors add new features to a class hierarchy without changing it. Iterators allow sequential access to elements of an object without changing it, allowing augmentation of it before iteration. Moving forward, I plan to use decorators, as they allow me to modify existing codebases, which means I can reuse other people's code without modifying it too much. Additionally, I have already used the iteration pattern to simplify my classes, only giving access to the required information when iterating. This LinkedIn Learning course has been useful in developing my coding skills.

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Linked in Learning

Python: Design Patterns

Course completed by Aaron Fletcher
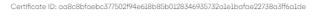Jun 23, 2024 at 08:35PM UTC · 2 hours 12 minutes

Top skills covered

(Design Patterns) (Python (Programming Language))

Head of Global Content, Learning

Certificate ID: aa8c8bfaebc377502f94e618b85b0128346935732a1e1bafae22738a3ff6a1de

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications

Training Course 6 - 2 hours

Title	Machine Learning and AI Foundations: Classification Modeling
Provider	LinkedIn Learning
Duration	2 Hours
Dates	29/06/2024
Narrative	This was a 2-hour online course linked to learning, which covered the basics of classification modelling. I chose this course to ensure that my understanding of classification topics matches that of other sources. The topics covered in the learning were: Classification Strategies Model Evaluation ● Confusion Matrics ● Lift Charts ● Gains Charts Classification Algorithms ● Stepwise Discriminant ● Logistic Regression ● Decision Trees ● KNN ● Linear SVMs ● Neural Networks ● Baysian Networks In honesty, this course was poorly given. While it provided little information that I was not aware of (which would mean that I should have selected a better course) by the time that the course was outlining its core teachings, about ¼ of the training course had been completed. For future training, I would ensure that I understand exactly what is being taught and ensure that I am the target audience. Additionally, some of the questions given in the chapter quiz were just plain incorrect (Groups other than healthcare professionals and social scientists can use linear regression, for example!).

UKRI Centre for Doctoral Training in
Speech and Language Technologies
and their Applications