COMP6713 – 2024 T1

Summary of Course Material

This document is a collection of reference material from the slides and the lecture content, and has been put together to help you with the preparation for the final exam. The lecture slides and tutorial notebooks are the primary material. Note the FAQs at the end of the document.

Hope you find this document useful. Good luck!

Dr. Aditya Joshi
Aditya.joshi@unsw.edu.au

Week 1: Introduction

Techniques: spacy, NLTK, HuggingFace pipelines, etc.

Chapter 1 of the Bhattacharyya-Joshi textbook.

	Key Idea	Demos
NLP Today & Yesterday	NLP has fascinated computer science for a long time.	Fundamental NLP tasks using NLTK
Ambiguity	Interaction between data, probability and ambiguity resolution	Text matching using spaCy
The three generations	Three-generational view of NLP	Open-source NLP models using HuggingFace
Considerations	NLP is far from solved. Hallucination, privacy, biases, etc.	Emerging NLP tools: Perplexity.ai

Week 2: Representation Learning

Techniques: Word2Vec, GloVe, probabilistic language modelling, etc.

Chapter 2 of the Jurafsky-Martin textbook:
https://web.stanford.edu/~jurafsky/slp3/3.pdf

Chapter 2 of the Bhattacharyya-Joshi textbook.

Word2Vec: Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.
GloVe: Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Part	Key Idea	Demos
Representation matters	One-hot vectors and their limitations	Vectorizer
Word2vec & GloVe	Word representation using context prediction or co-occurrence estimation	Word2Vec using gensim
Probabilistic language modeling	Language generation as conditional probability; smoothing helps.	Probabilistic language modeling using NLTK primitives
Sequential neural language modeling	RNNs/LSTMs can help mitigate the problem in probabilistic language modeling. However, linear structures limit the capability of the models.	Simple LSTM-based language model using Keras

Week 3: Transformer
Techniques: Attention, Transformer.

Chapter 2 of Bhattacharyya-Joshi textbook.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Attention: Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate.". ICLR 2015.

Transformer pseudocode: Phuong et al., Formal Algorithms for Transformers, https://arxiv.org/abs/2207.0923

Annotated Transformer: https://jalammar.github.io/illustrated-transformer/

Part	Key Idea	Demos
Why Attention?	Intuition behind attention	Pytorch tutorial
Attention in recurrent language modeling	Additive attention in recurrent LM and limitations	Attention Visualization
Encoders & decoders	Scaled dot-product attention -> Multi-head attention -> Self-attention, masked attention -> Encoders and decoders. Pseudocode + Animation.	Transformer tutorial
Tokenization and positional encoding	BPE algorithm, positional encoding, language model head	Transformer tutorial continued. BPE Tokenizer. Positional encoding Visualization
Transformer: Looking Back	Pseudocode; components; impact of Transformer.	

Week 4: Transformer-based Language Models
Topics: Encoder models, decoder models, LoRA.

Chapters 10 and 11 of the Jurafsky-Martin textbook.

Chapter 2 of the Bhattacharyya-Joshi textbook.

BERT: Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL. https://jalammar.github.io/illustrated-bert/
GPT: Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
LoRA: Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." ICLR. 2021.

Part	Key Idea	Demos
Derivatives of Transformer	Encoder and decoder models	
Encoder models	BERT pre-training and fine-tuning; Variants of BERT. "BERT is a vacuum cleaner"	BERT
Decoder models	GPT; prompting	GPT prompting
Fine-tuning methods	PEFT, Prompt/prefix-tuning, LoRA	Simple LoRA, OPT Fine-tuning using PEFT
Datasets & Libraries	NLP benchmarks and LangChain	LangChain

Week 5: Sentiment Analysis
Techniques: Lexical Resources, BERT fine-tuning, Chain-of-thought Prompting, Prompt Tuning, etc.

Chapter 8 of the Bhattacharyya-Joshi textbook.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In EMNLP 2013.

Baccianella, S., Esuli, A., & Sebastiani, F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In LREC 2010.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." EMNL P 2002.

Wang, Yequan, et al. "Attention-based LSTM for aspect-level sentiment classification." *Proceedings of EMNLP.*

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In ACL.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In EMNLP.

Additional Material:
Kanojia, Diptesh, and Aditya Joshi. "Applications and Challenges of Sentiment Analysis in Real-life Scenarios." Book Chapter, Computational Intelligence Applications for Text and Sentiment Data Analysis", Elsevier, 2023.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint Modelling of Emotion and Abusive Language Detection. In *ACL*.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, Guandong Xu, 'Leveraging Sentiment Distributions to Distinguish Figurative From Literal Health Reports on Twitter', TheWebConf (Ex-WWW) 2020.

Kertkeidkachorn, Natthawut, and Kiyoaki Shirai. "Sentiment Analysis using the Relationship between Users and Products." Findings of the Association for Computational Linguistics: ACL 2023. 2023.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In EMNLP 2020.

Part	Key Concepts	Demos
An umbrella term	Several sub-tasks, challenges and applications	-
Sentiment Lexicons and Datasets	SentiWordNet, dataset creation, annotation strategies	SentiWordNet
Using rules and features	Rules, feature engineering, embeddings as features	SA with unigrams and SVM
Pre-decoder SA	Linear chain models & BERT fine-tuning, extensions via external networks, multi-task learning, etc.	BERT fine-tuning
Decoder SA	CoT prompting, prompt search, prompt tuning	Prompt tuning

Module 7: POS Tagging & NER
Techniques: HMM, CRF, BiLSTM+CRF, etc.

Chapter 3 of the Bhattacharyya, Joshi textbook.

Chapter 8 & "A" of the Jurafsky-Martin textbook.

HMM-based tagging: Kupiec, Julian. "Robust part-of-speech tagging using a hidden Markov model." Computer speech & language 6.3 (1992): 225-242.

CRF-based tagging: Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." *In ICML*.

Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." arXiv preprint arXiv:1508.01991 (2015).
Lample, Guillaume, et al. "Neural Architectures for Named Entity Recognition." Proceedings of NAACL-HLT. 2016.

Ma, Xuezhe, and Eduard Hovy. "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF." ACL 2016.

Additional Material:

Christopher Manning. 2011. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In Proceedings of CICLING.

Owoputi, Olutobi, et al. "Improved part-of-speech tagging for online conversational text with word clusters." Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies. 2013.

Wang, Yu, et al. "Nested named entity recognition: a survey." ACM Transactions on Knowledge Discovery from Data (TKDD) 16.6 (2022): 1-29.

Nishida, Kosuke, Naoki Yoshinaga, and Kyosuke Nishida. "Self-Adaptive Named Entity Recognition by Retrieving Unstructured Knowledge." EACL 2023.

Part	Key Concepts	Demos
Two components of tagging	Tagsets for POS tagging and NER; BERT-based tagging; transition and observation	BERT
HMM + CRF	Generative and discriminative models	CRF
CRF->BiLSTM+CRF	Combining CRF with BiLSTM	BiLSTM+CRF
Specials cases of POS Tagging & NER	Emerging entities, domain-specific NER, nested NER	-

Module 8: Machine Translation
Techniques: Transformer decoding, Evaluation metrics, Unsupervised NMT, Instruction tuning, etc.

Chapter 7 of the Bhattacharyya-Joshi textbook.
Chapter 10 and 13 of the Jurafsky-Martin textbook.

BLEU: Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020.

https://uva-slpl.github.io/nlp2/resources/papers/CollinsIBM.pdf

Aharoni, Roee, Melvin Johnson, and Orhan Firat. "Massively Multilingual Neural Machine Translation." NAACL 2019.
Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research. 2020.

Wang, Yizhong, et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions." The 61st Annual Meeting Of The Association For Computational Linguistics. 2023.

Alves, Duarte M., et al. "Tower: An Open Multilingual Large Language Model for Translation-Related Tasks." arXiv preprint arXiv:2402.17733 (2024).

Artetxe, Mikel, et al. "Unsupervised Neural Machine Translation." *International Conference on Learning Representations*. 2018.

Additional Material:

Tay, Yi, et al. "Efficient transformers: A survey." *ACM Computing Surveys* 55.6 (2022): 1-28.
Maruf, Sameen, Fahimeh Saleh, and Gholamreza Haffari. "A survey on document-level neural machine translation: Methods and evaluation." *ACM Computing Surveys (CSUR)* 54.2 (2021): 1-36.
Ranathunga, Surangika, et al. "Neural machine translation for low-resource languages: A survey." *ACM Computing Surveys* 55.11 (2023): 1-37.
Apidianaki, Marianna. "From word types to tokens and back: A survey of approaches to word meaning representation and interpretation." Computational Linguistics 49.2 (2023): 465-523.
Tang, Yuqing, et al. "Multilingual translation with extensible multilingual pretraining and finetuning." arXiv preprint arXiv:2008.00401 (2020).

Part	Key Concepts	Demos
Introduction	Terminology; history	Zero-shot MT
MT Evaluation	BLEU, METEOR, ROUGE, BLEURT	Libraries to compute the metrics
Statistical MT	Alignment & language models; IBM models	-
Transformer & MT	Transformer decoding: Greedy/sampling/beam search	OpenNMT tutorial Decoding strategies
LLM-based MT	Instruction tuning TowerLLM	Instruction tuning for MT
Special cases of MT	Pivot-based MT Unsupervised MT: Denoising & backtranslation Automatic post-editing	

Module 9: Summarization
Techniques: Graph-based sentence selection, pointer-generator networks, window attention, encoder-decoder models, etc.
Primary Source
Chapter 11 of Bhattacharyya, Joshi, 'Natural Language Processing', Wiley, 2023.

Lin, Hui, and Vincent Ng. "Abstractive summarization: A survey of the state of the art." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into text." *EMNLP 2004*.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." ACL 2020.

See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get To The Point: Summarization with Pointer-Generator Networks." ACL, 2017.

Ma, Congbo, et al. "Multi-document summarization via deep learning techniques: A survey." ACM Computing Surveys 55.5 (2022): 1-37.

Wang, Qiqi, et al. "Towards Legal Judgment Summarization: A Structure-Enhanced Approach." ECAI2023 (2023).

Liu, Yizhu, Qi Jia, and Kenny Zhu. "Length control in abstractive summarization by pretraining information selection." ACL 2022.

Additional Material:
Multi-document Summarization: Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document Summarization via Deep Learning Techniques: A Survey. ACM Comput. Surv. 55, 5, Article 102 (May 2023), 37 pages. https://doi.org/10.1145/3529754

Dialogue Summarization: Zhao, Lulu, et al. "Domain-Oriented Prefix-Tuning: Towards Efficient and Generalizable Fine-tuning for Zero-Shot Dialogue Summarization." NAACL. 2022.

Part	Key Concepts	Demos
What and why	Terminology: Abstractive/Extractive; What is a good summary?	SparkNLP
Extractive summarization	Graph-based and classification-based methods	TextRank
Abstractive summarization	Pointer-generator networks, Denoising using encoder-decoder models (BART); windowed attention in Longformer	Longformer
Special cases of summarization	Multi-document summarization; domain-specific summarization; length-specific summaries (modified attention)	-

Module 10: Applications & Frontiers
Techniques: Retrieval-augmented generation, bias mitigation using prompt tuning, bias metrics, commonsense reasoning, etc.

Hallucination

Ji, Ziwei, et al. "Survey of hallucination in natural language generation." *ACM Computing Surveys* 55.12 (2023): 1-38.

Guo, Yue, Yi Yang, and Ahmed Abbasi. "Auto-debias: Debiasing masked language models with automated biased prompts." ACL. 2022.

Bias

Czarnowska, Paula, Yogarshi Vyas, and Kashif Shah. "Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics." Transactions of the Association for Computational Linguistics 9 (2021): 1249-1267.

William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task Agnostic Dialect Adapters for English. In Findings of the Association for Computational Linguistics: ACL 2023, pages 813–824, Toronto, Canada. Association for Computational Linguistics.

Guo, Yue, Yi Yang, and Ahmed Abbasi. "Auto-debias: Debiasing masked language models with automated biased prompts." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Reasoning

Storks, Shane, Qiaozi Gao, and Joyce Y. Chai. "Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches." arXiv preprint arXiv:1904.01172 (2019): 1-60.

Rajani, Nazneen Fatema, et al. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

FAQs:

1) Where can I find lecture recordings?

Answer: On Echo360 (accessed via Moodle)

2) Where can I find lecture handouts?

Answer: On webcms. Course work > Lectures

3) Will the final exam involve writing code?

Answer: You will not write a program. However, you may be required to fill a blank with a line of code, correct or explain a code.

4) Am I allowed to use ChatGPT or similar tools during the final exam?

Answer: No.

5) Am I allowed to sit with a textbook, notebook, printouts or a device showing lecture slides during the final exam?

Answer: No.

6) Will any additional information be collected from the final exam?

Answer: We will be using any login information that is provided by Moodle by default before we begin evaluating the final exam.

7) Can I use a UNSW-approved calculator during the exam?

Answer: Yes.

8) Can I seek help of an AI-based tool to answer any question in the final exam?

Answer: No.

9) Can I seek help of an human or group of humans to answer any question in the final exam?

Answer: No.