TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
THAPATHALI CAMPUS

Presentation on

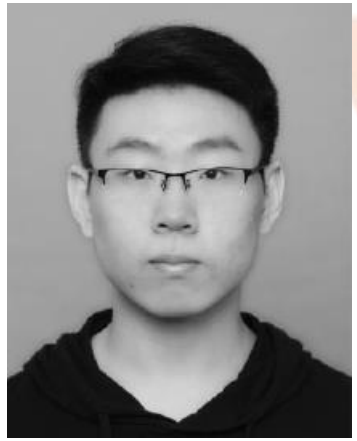
Paradigm Shift in Natural Language Processing

Presented By:

Bibat Thokar

THA078MSIISE02

PARADIGM SHIFT IN NATURAL LANGUAGE PROCESSING



Xiang-Yang Liu

Tian-Xiang Sun

Xi-Peng Qiu

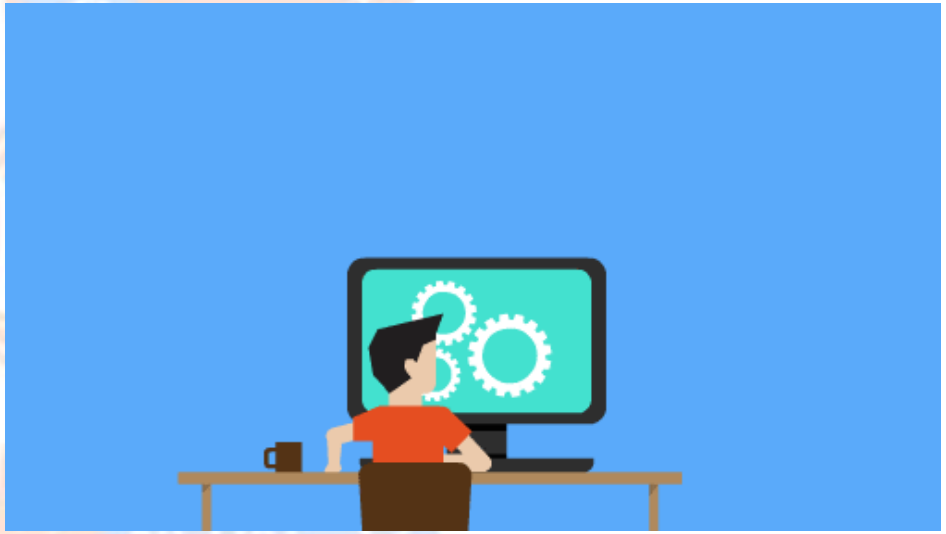
Xuan-Jing Huang

OUTLINES

- Motivation
- Introduction
- Objectives
- Theoretical Background
- Paradigm shift in NLP tasks
- Potential unified paradigms in NLP
- Advantage
- Conclusion
- References

MOTIVATION

- With the rapid progress of pre-trained language models, a rising trend of paradigm shift is solving NLP task by reformulating the task.
- The paradigm shift is becoming a promising way to improve model performance.
- Some of these paradigms have great potential to unify a large number of NLP tasks to build a single model to handle diverse tasks.
- The paper has reviewed phenomenon of paradigm shifts in recent years highlighting the potentials to solve different NLP tasks.



INTRODUCTION

- Paradigm is a general modeling framework or a distinct set of methodologies to solve a class of tasks.
- Different paradigms usually require different formats of input and output, and highly depend on the annotation of the tasks.
- Recent research has shown that some models under paradigms generalize well on tasks with other paradigms.
- Such models under some paradigms typically first convert the form of the dataset to the form required by the new paradigm and solves the task.
- After the emergence of pre-trained language models (PTM), some paradigms have shown great potential to unify diverse NLP tasks.

OBJECTIVE

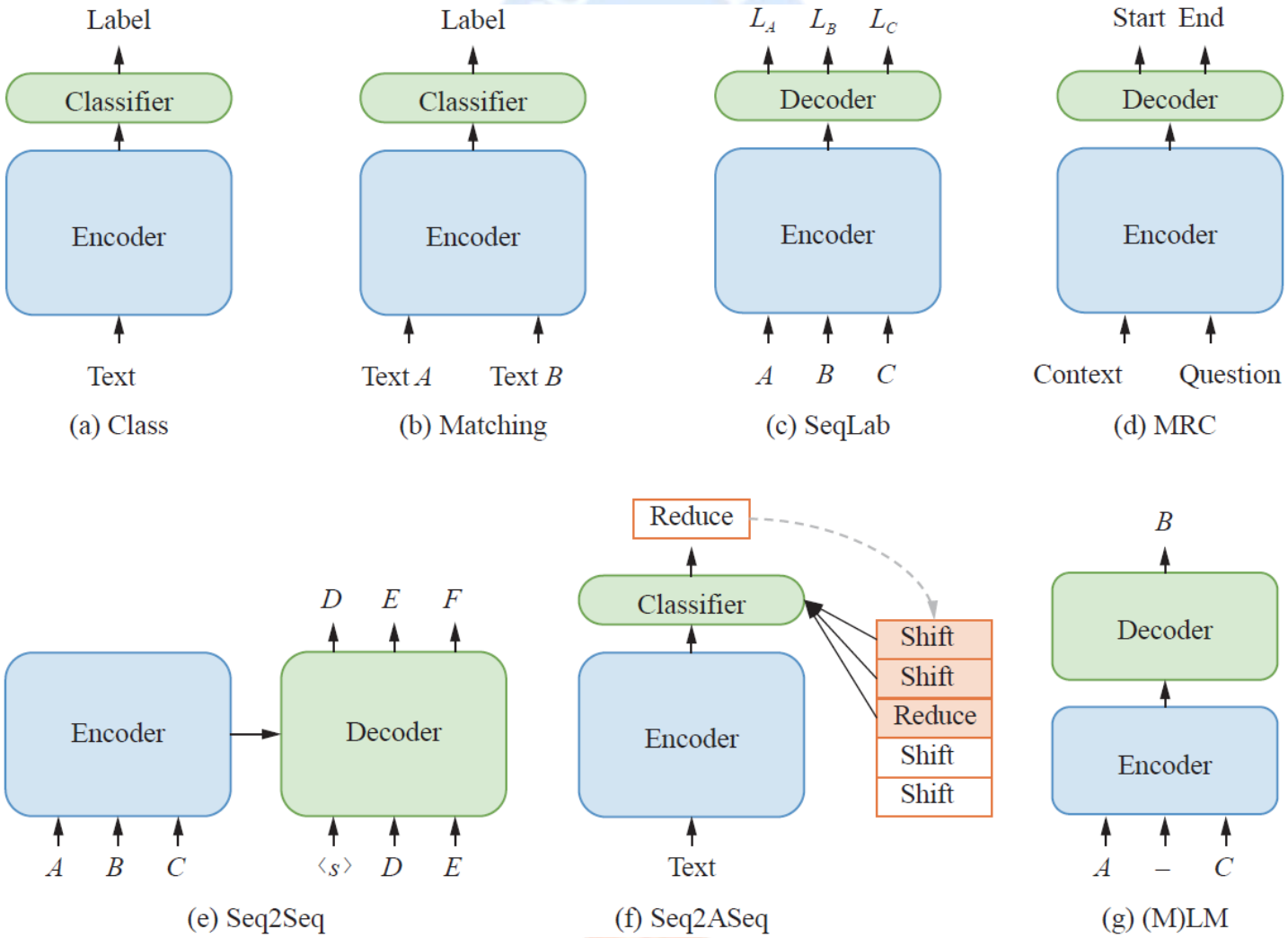
- Despite the success of these paradigm, the shifts scattering in various NLP tasks have not been systematically reviewed and analyzed.
- The paper mainly targets to make an attempt for summarizing recent advances and trends in NLP paradigm of research known as paradigm shift or paradigm transfer.

THEORETICAL BACKGROUND – [1]
PARADIGMS IN NLP

- A paradigm is the general modeling framework to fit some datasets D = $(X_i, Yi)_{i=1}^{N}$ i.e. tasks with a specific format of data structure of X and Y.
- A task can be solved by multiple paradigms by transforming it into different formats.
- Paradigm can be used to solve multiple tasks that can be formulated in a format and can be instantiated by a class of models with similar architectures.
- In this paper, the following seven paradigms that are widely used in NLP tasks are considered:

- Classification or class
- Matching
- Sequence Labeling (SeqLab)
- Masked Language modeling ((M)LM)
- Machine Reading Comprehension(MRC)
- Sequence to Sequence (Seq2Seq)
- Sequence to Action Sequence(Seq2ASeq)

SEVEN PARADIGMS IN NLP



THEORETICAL BACKGROUND-[2]
CLASSIFICATION (CLASS)

- Text classification has designating predefined labels for text which can be used for sentiment analysis, topic classification, spam detection etc.
- In deep learning, text classification is done by feeding the input text into a deep neural-based encoder.
- Encoder extract the task-specific feature which is then fed into a shallow classifier to predict the label. Mathematically,

  Y = Cls(Enc(X)), where Y can be one-hot or multi-hot in which is called multi-label classification.
- Enc(.) can be instantiated as convolutional networks, recurrent networks, or transformers.
- Cls(.) is usually implemented as a simple multi-layer perceptron following a pooling layer.

THEORETICAL BACKGROUND-[3]
MATCHING

- Text matching is a paradigm to predict the semantic relevance of two texts their fine-grained interactions
- It can be formulated as:

  $Y = Cls(Enc(Xa;Xb))$ where X_a and X_b are two texts to be predicted.
- Y is semantic similarity between the two texts and can be discrete or continuous.
- It is used in information retrieval, natural language inference (NLI), question answering and dialogue systems.

THEORETICAL BACKGROUND-[4]
SEQUENCE LABELING (SeqLab)

- Sequence labeling is used to model part-of-speech (POS) tagging, name entity recognition (NER), and text chunking.
- Conventional neural based sequence labeling models are consists of an encoder to capture the contextualized feature for each token in the sequence, and a decoder to take in the features and predict the labels.
- Mathematically,

  $y_1,\ldots, y_n = \text{Dec}(\text{Enc}(x_1,\ldots, x_n))$, where $y_1,\ldots, y_n$ are the corresponding labels of $x_1,\ldots, x_n$
- Enc(.) can be instantiated as a recurrent network or a transformer encoder and Dec(.) is usually implemented as conditional random fields (CRF).

THEORETICAL BACKGROUND-[5]
MACHINE READING COMPREHENSION (MRC)

- MRC extracts contiguous token sequences (spans) from the input sequence conditioned on a given question.
- Mathematically,

  $y_k, \ldots, y_{k+l} = \text{Dec}(\text{Enc}(X_p, X_q))$, where X_p and X_q denote passage (or context) and query, and $y_k, \ldots, y_{k+l}$ is a span from X_p or X_q.
- Typically, Dec(.) is implemented as two classifiers, one for predicting the starting position and another for predicting the ending position.

THEORETICAL BACKGROUND-[6]
SEQUENCE-TO-SEQUENCE (Seq2seq)

- Seq2Seq generates target language or response conditioned on an input source language or user query.
- It is typically implemented by an encoder-decoder framework as:

  $y_1,\ldots, y_m = \text{Dec}(\text{Enc}(x_1, \ldots, x_n))$

- It differs from SeqLab in such way that the lengths of the input and output are not necessarily the same.
- The decoder in Seq2Seq is usually more complicated and takes the previous output as input at each step when inference or the ground truth when training.

THEORETICAL BACKGROUND-[7]
SEQUENCE-TO-ACTION-SEQUENCE (Seq2ASeq)

- Seq2ASeq aims to predict an action sequence (i.e. transition sequence) from some initial configuration to a terminal configuration.
- The predicted action sequence should encode some legal structure, such as a dependency tree.
- The instances of the Seq2ASeq paradigm are usually called transition-based models.
- Mathematically,
  - $A = Cls(Enc(X), C)$ where $A = a_1, \ldots, a_m$ is a sequence of actions, and $C = c_0, \ldots, c_{m-1}$ is a sequence of configurations.
- At each time step, the model predicts an action a_t based on the input text and the current configuration c_{t-1}.

THEORETICAL BACKGROUND-[8]
MASKED LANGUAGE MODELING M(LM)

- LM estimates the probability of a given sequence of words occurring in a sentence.
- Masked language modeling M(LM) is adopted as training objectives to pre-train models on a large-scale unlabeled corpus which is based on auto-encoding.
- Mathematically,

  $x_k = Dec(x_1, \ldots, x_{k-1})$, where Dec(.) can be any auto-regressive model such as recurrent networks and transformer decoder.

- MLM can be formulated as:

  $\bar{x} = Dec(Enc(\tilde{x}))$, where $\tilde{x}$ is a corrupted version by replacing a portion of x, and $\bar{x}$ denotes the masked tokens to be predicted.

  Dec(.) can be implemented as a simple classifier as in bidirectional encoder.

PARADIGM SHIFT IN NLP TASKS

- The paradigm shifts that occur in different NLP tasks are:
  - Text Classification
  - Natural Language Inference
  - Named Entity Recognition
  - Aspect-based Sentiment Analysis
  - Relation Exaction
  - Text Summarization
  - Parsing.

PARADIGM SHIFT IN NLP TASKS

	Task	Original paradigm		Shifted paradigm			
TC	Paradigm	Class	⇒	Matching	Seq2Seq	(M)LM	
	Input	$\mathcal{X}$		$\mathcal{X}, \mathcal{L}$	$\mathcal{X}$	$f_{prompt}(\mathcal{X})$	
	Output	$\mathcal{Y}$		$\mathcal{Y} \in \{0,1\}$	$y_1, \cdots, y_m$	$g(\mathcal{Y})$	
	Example	[3]		[36]	[37]	[7]	
NLI	Paradigm	Matching	⇒	Class	Seq2Seq	(M)LM	
	Input	$\mathcal{X}_a, \mathcal{X}_b$		$\mathcal{X}_a \oplus \mathcal{X}_b$	$f_{prompt}(\mathcal{X}_a, \mathcal{X}_b)$	$f_{prompt}(\mathcal{X}_a, \mathcal{X}_b)$	
	Output	$\mathcal{Y}$		$\mathcal{Y}$	$\mathcal{Y}$	$g(\mathcal{Y})$	
	Example	[18]		[3]	[38]	[7]	
NER	Paradigm	SeqLab	⇒	Class	MRC	Seq2Seq	(M)LM
	Input	$x_1, \cdots, x_n$		$\mathcal{X}_{span}$	$\mathcal{X}, \mathcal{Q}_y$	$\mathcal{X}$	$\mathcal{X}$
	Output	$y_1, \cdots, y_n$		$\mathcal{Y}$	$\mathcal{X}_{span}$	$(\mathcal{X}_{ent_i}, \mathcal{Y}_{ent_i})_{i=1}^m$	$g(\mathcal{Y})$
	Example	[19]		[39]	[1]	[2]	[40]
ABSA	Paradigm	Class	⇒	Matching	MRC	Seq2Seq	(M)LM
	Input	$\mathcal{X}_{asp}$		$\mathcal{X}, \mathcal{S}_{aux}$	$\mathcal{X}, \mathcal{Q}_{asp}, \mathcal{Q}_{opin}, \mathcal{Q}_{sent}$	$\mathcal{X}$	$f_{prompt}(\mathcal{X})$
	Output	$\mathcal{Y}$		$\mathcal{Y}$	$\mathcal{X}_{asp}, \mathcal{X}_{opin}, \mathcal{Y}_{sent}$	$(\mathcal{X}_{asp_i}, \mathcal{X}_{opin_i}, \mathcal{Y}_{sent_i})_{i=1}^m$	$g(\mathcal{Y})$
	Example	[41]		[42]	[43]	[44]	[45]
RE	Paradigm	Class	⇒	MRC	Seq2Seq	(M)LM	
	Input	$\mathcal{X}$		$\mathcal{X}, \mathcal{Q}_y$	$\mathcal{X}$	$f_{prompt}(\mathcal{X})$	
	Output	$\mathcal{Y}$		$\mathcal{X}_{ent}$	$(\mathcal{Y}_i, \mathcal{X}_{sub_i}, \mathcal{X}_{obj_j})_{i=1}^m$	$g(\mathcal{Y})$	
	Example	[46]		[47]	[48]	[49]	
Summ	Paradigm	SeqLab / Seq2Seq	⇒	Matching	(M)LM		
	Input	$\mathcal{X}_1, \cdots, \mathcal{X}_n$ / $\mathcal{X}, \mathcal{Q}_{summ}$		$(\mathcal{X}, \mathcal{S}_{cand_i})_{i=1}^n$	$\mathcal{X}$, Keywords/Prompt		
	Output	$\mathcal{Y}_1, \cdots, \mathcal{Y}_n \in \{0,1\}^n$ / $\mathcal{Y}$		$\hat{\mathcal{S}}_{cand}$	$\mathcal{Y}$		
	Example	[38, 50]		[51]	[52]		
Parsing	Paradigm	Seq2ASeq	⇒	(M)LM	SeqLab	MRC	Seq2Seq
	Input	$(\mathcal{X}, \mathcal{C}_t)_{t=0}^{m-1}$		$(\mathcal{X}, \mathcal{Y}_i)_{i=1}^k$	$x_1, \cdots, x_n$	$\mathcal{X}, \mathcal{Q}_{child}$	$\mathcal{X}$
	Output	$\mathcal{A} = a_1, \cdots, a_m$		$\hat{\mathcal{Y}}$	$g(y_1, \cdots, y_n)$	$\mathcal{X}_{parent}$	$g(y_1, \cdots, y_m)$
	Example	[28]		[53]	[54]	[55]	[56]

TEXT CLASSIFICATION

- Conventional text classification tasks can be well solved by the Class paradigm but its variants can be challenging in which case Class may be sub-optimal.
- It is proposed to adopt the Seq2Seq paradigm to better capture interactions between labels for multi-label classification tasks.
- Matching paradigm is adopted to predict whether the pair-wise input (X, L_y) is matched, where X is the original text and L_y is the label description for class.
- With the rise of pre-trained language models (LMs), text classification tasks can also be solved in the (M)LM paradigm.
- Gap between LM pre-training and fine-tuning is narrowed by reformulating a text classification task into a masked language modeling task resulting improved performance in limited training data.

NATURAL LANGUAGE INFERENCE (NLI)

- In NLI, the two input texts (X_a, X_b) are encoded and interact with each other followed by a classifier to predict the relationship between them.
- NLI tasks can be simply solved in the Class paradigm by concatenating the two texts as one with the help of powerful encoders such as Bidirectional Encoder Representations from Transformers (BERT).
- In case of few-shot learning, NLI tasks can also be formulated in the (M)LM paradigm by modifying the input.
- The unfilled token [MASK] can be predicted by the M(LM) head as Yes/No/Maybe, corresponding to Entailment/Contradiction/Neutral.

NAMED ENTITY RECOGNITION (NER)

- NER can be categorized into three subtasks: flat, nested, and discontinuous NER.
- Flat NER and nested NER are solved with the Class paradigm but it introduced the span overlapping problem.
- A heuristic decoding method is adopted to handle by keeping the span with the highest prediction probability.
- Flat NER and nested NER are formulated as an MRC task by reconstructing each sample into a triplet (X, Q_y, X_{span}) as given context, question, and answer respectively.
- Unified model based on the Seq2Seq paradigm is used to solve the three types of NER subtasks which achieved state-of-the-art performance on various datasets
- The input of the Seq2Seq paradigm is the original text, while the output is a sequence of span-entity pairs.

ASPECT-BASED SENTIMENT ANALYSIS (ABSA)

- A fine-grained sentiment analysis task has seven subtasks.
- MRC paradigm is adopted to handle all of the ABSA subtasks where two queries are constructed to sequentially extract the aspect terms with their corresponding polarities and opinion terms.
- ABSA subtasks are solved with the Seq2Seq paradigm along with equipped with BART as the backbone.
- In this approach the original label of a subtask is converted into a sequence of tokens, which is used as a target to train a Seq2Seq model.
- The ABSA subtasks can be formulated with the M(LM) paradigm where a consistency prompt, and a polarity prompt are constructed for the input text ‘X’, and the aspect ‘A’ along with opinion ‘O’ of interest.

RELATION EXTRACTION (RE)

- RE has two main subtasks as relation prediction (i.e. predicting the relationship of two given entities ‘s’ and ‘o’ conditioned on their context) and triplet extraction (i.e. extracting the triplet (s, r, o) from the input text).
- Relation prediction subtask is solved mainly with the Classification paradigm
- Triplet extraction subtask is solved in the pipeline style which first uses the SeqLab paradigm to extract the entities and then uses the paradigm to predict the relationship between the entities.
- RE task through MRC paradigm by generating relation-specific questions has a potential for zeroshot generalization to unseen relation types which extract entities and relations.
- M(LM) task using logic rules encodes prior knowledge of entities and relations into prompts and prompt tuning with rules (PTR), achieve state-of-the-art performance on multiple RE datasets.

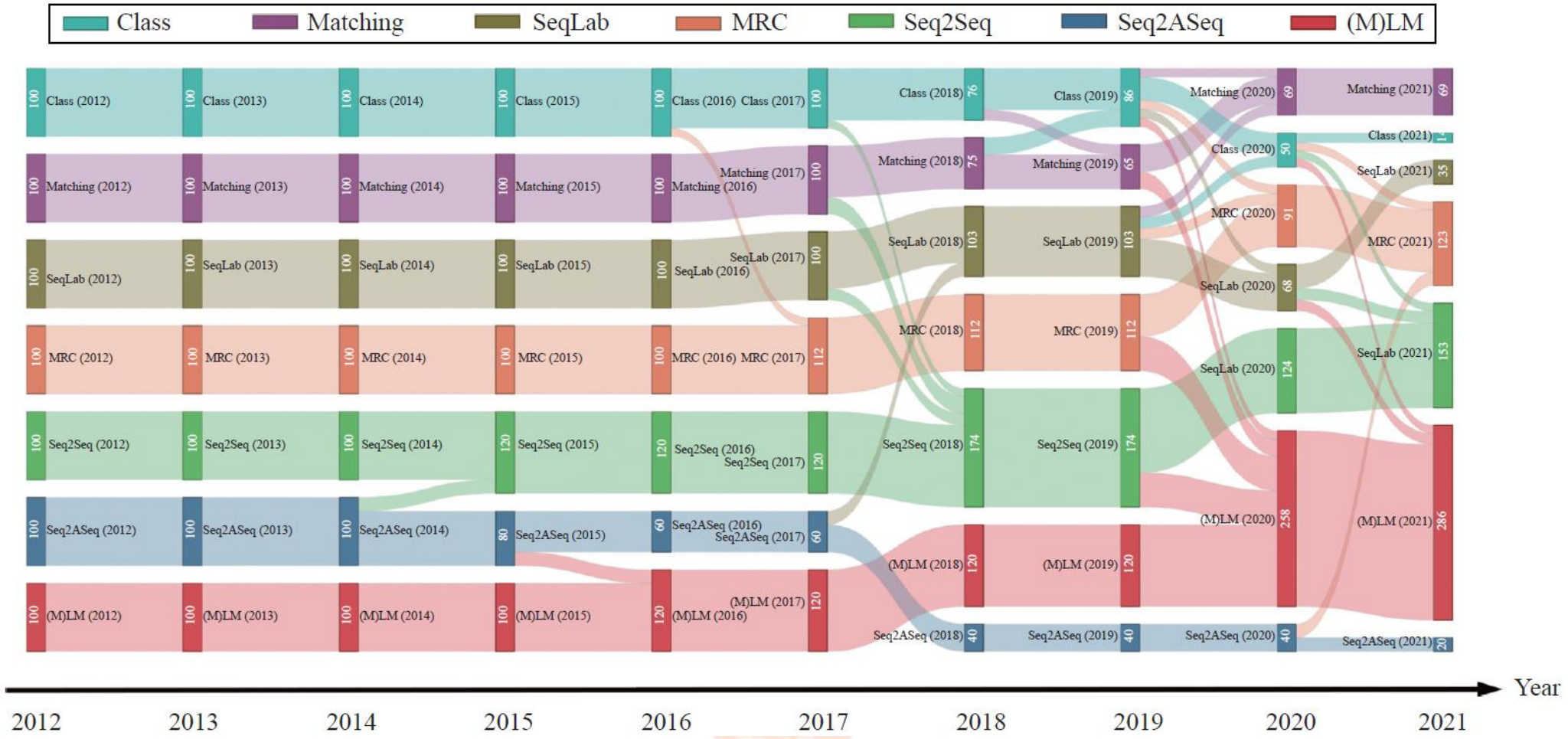
TEXT SUMMARIZATION

- Aims to generate a concise and informative summary of large texts.
- Two approaches to solve as extractive summarization and abstractive summarization.
- Extractive summarization approach extracts the clauses of the original text to form the final summary which usually lies in the SeqLab paradigm.
- Abstractive summarization approaches usually adopt the Seq2Seq paradigm to directly generate a summary conditioned on the original text.
- Extractive summarization task is solved using Matching paradigm by matching the semantics of the original text and each candidate summary finding with the highest matching score.
- The (M)LM paradigm pre-train a BART-style model directly on large-scale structured HTML web pages which is able to perform zero-shot text summarization by predicting the <title> element with the <body> of the document.

PARSING

- Parsing family of tasks is to derive a structured syntactic or semantic representation from a natural language utterance.
- Two commonly used approaches for parsing are transition-based methods and graph-based methods.
- Transition-based methods lie in the Seq2ASeq paradigm, and graph-based methods lie in the Class paradigm.
- Parsing can be solved in the Seq2Seq paradigm, the SeqLab paradigm, and the (M)LM paradigm by linearizing the target tree-structure to a sequence.
- MRC paradigm achieves state-of-the-art performance on dependency parsing tasks across various languages by extracting the parent span on the original sentence as context and the child span as the question.

POTENTIAL UNIFIED PARADIGMS IN NLP [SANKEY DIAGRAM]



POTENTIAL UNIFIED PARADIGMS IN NLP

- The frequency of paradigm shifts has been increasing especially after the emergence of pre-trained language models (PTMs).
- Hence to fully utilize the power of these PTMs, a better way is to reformulate various NLP tasks into the paradigms where PTMs are good at.
- More and more NLP tasks have shifted from traditional paradigms such as Class, SeqLab and Seq2ASeq to paradigms that are more general and flexible (M)LM, Matching, MRC and Seq2Seq.
- General paradigms that have the potential to unify diverse NLP tasks as (M)LM, Matching, MRC, and Seq2Seq have the different design challenges which will be discussed funther.

MASKED LANGUAGE MODELING (M)LM

- Reformulating downstream tasks into an (M)LM task is a natural way to utilize the pre-trained LMs.
- Original input is modified with a pre-defined or learned prompt with some unfilled slots, which can be filled by the pre-trained LMs
- The task labels can be derived from the filled tokens.
- For example, a movie review “I love this movie” can be modified by appending a prompt as “I love this movie. It was [MASK]”, in which [MASK] may be predicted as “fantastic ” by the LM. Then the word “fantastic ” can be labeled “positive ” by a verbalizer.
- Prompt-based learning has demonstrated great power in few-shot and even zero-shot settings by fully utilizing the pre-trained parameters of the (M)LM head instead of training a classification head from scratch.

SIDE NOTE – [PROMPT]

- Manually designed
  - Heuristic and non-intuitive, hand-crafted prompts that have achieved competitive performance on various few shot tasks.
- Mined from corpora
  - Prompts for relation extraction by mining sentences with the same subject and object in the corpus.
- Generated by paraphrasing
  - Use back translation to paraphrase the original prompt into multiple new prompts.
- Generated by another pre-trained language model
  - Prompts which is pre-trained to fill in missing spans in the input.
- Learned by gradient descent.
  - Prompts based on gradient guided search.

SIDE NOTE – [VERBALIZER]

- Manually designed.
  - Verbalizers which is not always intuitive for many tasks to manually design proper verbalizers.
- Automatically searched
  - Set of labelled data by minimizing some objective, such as negative log likelihood.
- Constructed and refined with knowledge base

SIDE NOTE – [PARAMETER-EFFICIENT PROMPT TUNING]

- Due to the parameter efficiency, prompt-based tuning is a promising technique for the deployment of large-scale pre-trained LMs.
- In traditional fine-tuning, the server has to maintain a task-specific copy of the entire pre-trained LM for each downstream task, and the inference has to be performed in separate batches.
- In prompt-based tuning, only a single pre-trained LM is required, and different tasks can be performed by modifying the inputs with task-specific prompts.

MATCHING

- Matching i.e. textual entailment is the task of predicting two given sentences, premise and hypothesis whether the premise entails the hypothesis, contradicts the hypothesis, or neither.
- Almost all text classification tasks can be reformulated as a textual entailment task
- For example, a labeled movie review {x : I love this movie, y: positive} can be modified as {x : I love this movie This is a great movie, y: entailment}.
- LMs that are fine-tuned on some large-scale annotates entailment datasets such as the multi-genre natural language inference (MultiNLI) dataset.
- To obtain the entailment model in a supervised fashion, the next sentence prediction head of BERT can be used on various zero-shot tasks without training on any supervised entailment data.

SIDE NOTE – [DOMAIN ADAPTATION]

- The entailment model may be biased to the source domain, resulting in poor generalization to target domains.
- To mitigate the domain difference between the source task and the target task, the cross-task nearest neighbor module was proposed.
- The module matches instance representations and class representations in the source domain and the target domain.

SIDE NOTE – [LABEL DESCRIPTIONS]

- For single-sentence classification tasks, the label descriptions for each class are required to be concatenated with the input text to be predicted by the entailment model.
- Label descriptions can be regarded as a kind of prompt to trigger the entailment model.
- Hand-crafted label descriptions with minimum domain knowledge can achieve state-of-the-art performance on various few-shot tasks.
- Nevertheless, human-written label descriptions can be sub-optimal, reinforcement learning is utilized to generate label descriptions.

SIDE NOTE – [COMPARISON WITH PROMPT-BASED LEARNING]

- In both (M)LM and Matching, the goal is to reformulate the downstream tasks into the pre-training task i.e. language modeling or entailment by modifying the input text with some templates.
- In prompt-based learning, the prediction is conducted by the pre-trained (M)LM head on the [MASK] token, while in matching-based learning, the prediction is conducted by the pre-trained classifier on the [CLS] token.
- In prompt-based learning, the output prediction is over the vocabulary, such that a verbalizer is required to map the predicted word in vocabulary into a task label.
- Matching-based learning can construct pairwise augmented data to perform contrastive learning, achieving a further improvement of few-shot performance.
- Matching-based learning can only be used in understanding tasks, while prompt-based learning can also be used for generation.

MACHINE LEARNING COMPREHENSION (MRC)

- MRC paradigm can be applied as long as the task input can be reformulated as context, question, and answer.
- Due to universality, decaNLP is proposed to unify ten NLP tasks but the answer may not appear in the context such as semantic parsing.
- Hence the framework is strictly not an MRC paradigm.

SIDE NOTE - [COMPARISON WITH PROMPT-BASED LEARNING]

- It is worth noticing that the designed question can be analogous to the prompt in (M)LM.
- Verbalizer is not necessary for MRC since the answer is a span in the context or question.
- Predictor i.e. (M)LM head in the prompt-based learning can be replaced by a start/end classifier as in traditional MRC models.

SEQ2SEQ

- Seq2Seq particularly suited for complicated tasks such as structured prediction which is compatible with other paradigms such as (M)LM, MRC.
- Sequential nature of auto-regressive fashion results in inherent latency at inference time.
- Hence more work is needed to develop efficient Seq2Seq models through non-autoregressive methods or other alternative techniques.

ADVANTAGE OF PARADIGM SHIFT

- Some of the paradigms have demonstrated the potential ability to formulate various NLP tasks into a unified framework.
- Advantages of a single unified model over multiple task-specific models are as follows:
- Data efficiency
  - Unified model has achieved considerable performance with much less labeled data.
- Generalization
  - Unified model can generalize to unseen tasks by formulating them into proper formats.
- Convenience
  - Unified models are easier and cheaper to deploy and serve, making them favorable as commercial black-box APIs.

CONCLUSION

- Prompt-based tuning can achieve considerable performance with less training data.
- Potential unified paradigms such as Matching, MRC and Seq2Seq are under- explored in pre-training because they require large scale annotated data.
- Matching requires less engineering, MRC is more interpretable, and Seq2Seq is more flexible to handle complicated tasks.
- These paradigms can achieve better performance than (M)LM by combining with self-supervised or further pre-training on annotated data.
- More attention is needed for the exploration of more powerful entailment, MRC, or Seq2Seq models through pre-training or other alternative techniques.

REFERENCES – [1]

- Paradigm shift in natural language processing, TX Sun, XY Liu, XP Qiu, XJ Huang - Machine Intelligence Research, 2022 – Springer
- X. Y. Li, J. R. Feng, Y. X. Meng, Q. H. Han, F. Wu, J. W. Li. A unified MRC framework for named entity recognition. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics
- H. Yan, T. Gui, J. Q. Dai, Q. P. Guo, Z. Zhang, X. P. Qiu. A unified generative framework for various NER subtasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL
- J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Minneapolis, USA

REFERENCES – [2]

- C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Q. Zhou, W. Li, P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research
- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. Language models are few-shot learners.
- X. P. Qiu, T. X. Sun, Y. G. Xu, Y. F. Shao, N. Dai, X. J. Huang. Pre-trained models for natural language processing: A survey. Science China Technological Sciences
- T. Schick, H. Schütze. Exploiting cloze-questions for few shot text classification and natural language inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics

REFERENCES – [3]

- T. Schick, H. Schütze. It's not just size that matters: Small language models are also few-shot learners. In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
- T. Y. Gao, A. Fisch, D. Q. Chen. Making pre-trained language models better few-shot learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing
- T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In Proceedings of Conference on Empirical Methods in Natural Language Processing
- X. L. Li, P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

REFERENCES – [4]

- X. Liu, Y. N. Zheng, Z. X. Du, M. Ding, Y. J. Qian, Z. L. Yang, J. Tang. GPT understands
- B. Lester, R. Al-Rfou, N. Constant. The power of scale for parameter-efficient prompt tuning
- T. X. Sun, Y. F. Shao, H. Qian, X. J. Huang, X. P. Qiu. Black-box tuning for language-model-as-a-service
- Y. Kim. Convolutional neural networks for sentence classification. In Proceedings of Conference on Empirical Methods in Natural Language Processing
- P. F. Liu, X. P. Qiu, X. J. Huang. Recurrent neural network for text classification with multi-task learning
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, \L. Kaiser, I. Polosukhin. Attention is all you need

REFERENCES – [5]

- Q. Chen, X. D. Zhu, Z. H. Ling, S. Wei, H. Jiang, D. Inkpen. Enhanced LSTM for natural language inference
- G. Paolini, B. Athiwaratkun, J. Krone, J. Ma, A. Achille, R. Anubhai, C. N. dos Santos, B. Xiang, S. Soatto. Structured prediction as translation between augmented natural languages.
- J. T. Gu, J. Bradbury, C. M. Xiong, V. O. K. Li, R. Socher. Non-autoregressive neural machine translation.
- W. Z. Qi, Y. Y. Gong, J. Jiao, Y. Yan, W. Z. Chen, D. Liu, K. W. Tang, H. Q. Li, J. S. Chen, R. F. Zhang, M. Zhou, N. Duan. BANG: Bridging autoregressive and non-autoregressive generation with large scale pretraining.
- M. Elbayad, J. T. Gu, E. Grave, M. Auli. Depth-adaptive transformer. In Proceedings of the 8th International Conference on Learning Representations


त्रिभुवन
विश्वविद्यालय
TRIBHUVAN UNIVERSITY ORGANISED 1956 A.D.
INCORPORATED 1959 A.D.
KATHMANDU
NEPAL

THANK YOU!!!