Graduate School of Fundamental Science and Engineering
Waseda University

# 博　士　論　文　概　要
# Doctoral Dissertation Synopsis

論　文　題　目
Dissertation Title

Towards Building a DIKW Pyramid for Conversational Systems

会話システムのためのDIKWピラミッド構築に向けて

申　請　者
(Applicant Name)
Junjie WANG
王　軍傑

Department of Computer Science and Communications Engineering, Research on Information Access

May, 2024

## 1 Introduction:

The emergence of the human-knowledge-AI loop represents a paradigm shift in how artificial intelligence (AI) interacts with human knowledge creation and utilization. This loop signifies the continuous cycle where humans generate knowledge, AI systems learn from this knowledge to serve human needs, and their outputs inspire further knowledge generation. However, integrating AI into this loop poses great challenges, particularly in understanding and processing human-centric information (Data, Information, Knowledge, Wisdom). This thesis investigates methodologies to effectively embed AI within this loop, focusing on enhancing conversational systems to understand, learn from, and contribute to human knowledge through the DIKW pyramid.

## 2 Challenges and Objectives:

The AI has brought transformative changes in various fields, including healthcare, education, and technology, by creating a dynamic human-knowledge-AI loop. This loop represents a continuous cycle of knowledge generation, learning, and application, where humans and conversational systems interact to enhance each other's capabilities. However, the integration of conversational systems within the human-knowledge-AI loop confronts two main challenges:

Complexity of Human-Centric Knowledge: Integrating AI within the human-knowledge-AI loop presents several challenges due to the necessity of developing distinct methodologies tailored to each layer of the DIKW (Data, Information, Knowledge, Wisdom) pyramid. The intrinsic complexity stems from the varying nature of information across these levels. For instance, at the data level, conversational systems must efficiently collect and process vast amounts of raw, unstructured facts. As we move to information, the challenge shifts towards organizing this data into meaningful constructs that provide context and relevance, such as uncertainty. At the knowledge layer, the focus is on synthesizing this information into actionable insights, requiring advanced reasoning and decision-making capabilities. Finally, at the wisdom level, AI systems must apply this knowledge with ethical insight and judgment, considering the broader implications of their outputs. The demand for unique approaches at each stage of the DIKW pyramid shows the challenge of developing AI systems capable of processing and understanding information in a way that mirrors human cognitive processes and ethical reasoning.

Practical Application Across Diverse Domains: The second challenge is the practical integration of these conversational systems into varied real-world domains, such as natural language processing (NLP) and multimodal information processing. Whether enhancing customer service, educational tools, or healthcare assistants, AI systems must demonstrate adaptability and the ability to provide reasonable responses.

In response to these challenges, this thesis sets two primary objectives:

Firstly, we aim to develop methodologies that enable conversational systems to effectively process and understand human-centric knowledge across the data, information, knowledge, and wisdom levels. This involves creating algorithms and frameworks capable of not only parsing and organizing vast amounts of data but also extracting meaningful information, understanding knowledge, and applying this knowledge with insight and ethical consideration. We seek to bridge the gap between human cognitive processes and AI capabilities, enabling conversational systems to engage in meaningful, context-aware, and insightful interactions.

Secondly, we aim to demonstrate the practical applications and benefits of these enhanced conversational systems

across various domains. This exploration into conversational systems' effectiveness spans straightforward query responses to advanced decision-making support. Within the realm of NLP, we collect the challenging benchmarks from a variety of tasks, including natural language inference, commonsense reasoning and coreference resolution. In terms of multimodal information processing, we explore several practical scenarios such as visual question answering, visual reasoning and visual entailment.

**3 Contributions:**

**Contribution 1: Learning multi-source data through trilinear model and workflow optimization**

**Contribution 2: Modeling multimodal uncertainty with the probability distribution encoder module**

**Contribution 3: Enhancing zero-shot learning by learning knowledge with UniMC framework**

**Contribution 4: Introducing wisdom into conversational systems through an ethical alignment process**

In response to the challenge of processing diverse layers of information, we develop a suite of new methods. These include an encoder designed for effective data integration, a flexible plugin for information processing, a comprehensive framework for knowledge acquisition and decision-making, and a decoupled design for the infusion of wisdom. Each method is specifically tailored to address the distinct characteristics of human-centric information, showcasing an effective synergy between structural design and the multifaceted dimensions of knowledge.

***Contribution 1. Learning multi-source data through trilinear model and workflow optimization***

For learning multi-source data, we introduce the MIRTT, a new end-to-end trilinear interaction model. The MIRTT enhances inter-modal and intra-modal interactions via a trilinear attention mechanism. To accommodate the complexities of free-form open-ended (FFOE) problems, we propose a streamlined two-stage workflow. Through our methodology, MIRTT has set new benchmarks, achieving state-of-the-art (SOTA) performances on various multiple-choice (MC) tasks, including the Visual7W telling and VQA1.0 MC tasks. Additionally, it has outperformed existing baseline models on the VQA2.0, TDIUC, and GQA datasets, demonstrating advancements in handling FFOE challenges.

***Contribution 2. Modeling multimodal uncertainty with the probability distribution encoder module***

To capture the multimodal uncertainty within the information, we introduce a new plug-and-play module named the Probability Distribution Encoder (PDE). Our proposed PDE enhances the effectiveness of various frameworks in multimodal understanding and generation tasks. To the best of our knowledge, this is the first integration of uncertainty learning into video captioning. Furthermore, we develop three distribution-based pre-training strategies, aimed at learning multimodal uncertainties in large-scale unlabeled datasets. We integrate these pre-training tasks and PDE into an end-to-end multimodal uncertainty-aware vision-language pre-training (MAP) framework. Empirical evaluations demonstrate that the MAP model achieves SOTA performance across multiple downstream tasks such as MSRVTT, MSCOCO, Flickr30K, VQA2.0, NVLR2 and SNLI-VE.

***Contribution 3. Enhancing zero-shot learning by learning knowledge with UniMC framework***

To address unseen problems using knowledge, we introduce an experience-driven framework, the unified multiple-choice model (UniMC). By transforming label-based NLP tasks into multiple-choice formats, we minimize the need for manual processing. Further, we incorporate option mask language modeling (O-MLM) and option prediction (OP) tasks, along with a multiple-choice tuning method. These strategies are effectively integrated within the UniMC framework to tackle unknown challenges. In experiments, UniMC achieves SOTA performance in both in-domain and out-of-domain tasks, such as achieving up to a 48% improvement on the Dbpedia dataset. We also develope the UniMC-Chinese model based on DeBERTa-v2, which surpasses human benchmarks in several tasks.

***Contribution 4. Introducing wisdom into conversational systems through an ethical alignment process***

To infuse human wisdom into conversational systems, we develop a ethical alignment process (EAP). This process is designed to be decoupled and integrated within existing conversational systems. We enhanced the evaluation of moral alignment through the reconstruction of the QA-ETHICS dataset. Moreover, we introduce the MP-ETHICS dataset, which assesses multi-perspective ethical capabilities. Additionally, we propose the ethical alignment language model (EALM), a new framework that achieves the SOTA results across multiple ethics benchmarks. For example, on the hard set of the ETHICS dataset, our EALM framework improve accuracy by a 23.2% over the existing top-performing models.

## 4 Organization of the Thesis

We organize the thesis as follows.

- In Chapter 1, we introduce the background, motivations and contributions.
- In Chapter 2, we outline recent works in the field, focusing on two primary areas: the DIKW pyramid and conversational systems.
- In Chapter 3, we discuss a common requirement in integrating various data sources into conversational systems: the challenge of merging multiple diverse sources. We present the proposed MIRTT model and the two-stage workflow. (Contribution 1)
- In Chapter 4, we aim for the conversational systems to mine information from the data, further understanding its inherent uncertainties and other high-level characteristics. We introduce a probability distribution encoder, which is a plugin-and-play module designed to model the uncertainties. Additionally, we develop new pre-training strategies focused on unlocking the uncertainties of unlabeled data. (Contribution 2)
- In Chapter 5, following our discussion on data and information, we explore the integration of knowledge, highlighting the need to emulate human decision-making behaviors, particularly in zero-shot scenarios. We discuss the proposed UniMC methods and their experimental outcomes. (Contribution 3)
- In Chapter 6, we explore integrating abstract human wisdom, especially ethics, into the conversational systems. We introduce a decoupled EAP and propose a EALM for addressing multidimensional ethical alignment problems. (Contribution 4)
- In Chapter 7, we consolidate all discoveries made throughout the thesis, presenting a comprehensive overview of our research outcomes.

# List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 王　軍傑　　seal or signature

Date Submitted(yyyy/mm/dd): 2024/04/18

| 種類別 (By Type) | 題名、 発表・発行掲載誌名、 発表・発行年月、 連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself) |
|---|---|
| Journal | ○ Junjie Wang, Yatai Ji, Yuxiang Zhang, Yanru Zhu, and Tetsuya Sakai. Modeling multimodal uncertainties via probability distribution encoders included vision-language models. IEEE Access, 12:420–434, 2024.<br><br>○ Junjie Wang, Ping Yang, Ruyi Gan, Yuxiang Zhang, Jiaxing Zhang, and Tetsuya Sakai. Zero-shot learners for natural language understanding via a unified multiple-choice perspective. IEEE Access, 11:142829–142845, 2023.<br><br>Yuxiang Zhang, Junjie Wang, Xinyu Zhu, Tetsuya Sakai, and Hayato Yamana. SSR: solving named entity recognition problems via a single-stream reasoner. To appear. ACM Transactions on Information Systems, 2024. |
| International Conference | ○ Junjie Wang, Yatai Ji, Jiaqi Sun, Yujiu Yang, and Tetsuya Sakai. MIRTT: learning multimodal interaction representations from trilinear transformers for visual question answering. In EMNLP (Findings), pages 2280–2292. Association for Computational Linguistics, 2021.<br><br>○ Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. MAP: multimodal uncertainty-aware vision-language pre-training model. In CVPR, pages 23262–23271. IEEE, 2023.<br><br>○ Ping Yang, Junjie Wang, Ruyi Gan, Xinyu Zhu, Lin Zhang, Ziwei Wu, Xinyu Gao, Jiaxing Zhang, and Tetsuya Sakai. Zero-shot learners for natural language understanding via a unified multiple choice perspective. In EMNLP, pages 7042–7055. Association for Computational Linguistics, 2022.<br><br>○ Yiyao Yu, Junjie Wang, Yuxiang Zhang, Lin Zhang, Yujiu Yang, and Tetsuya Sakai. Ealm: Introducing multidimensional ethical alignment in conversational information retrieval. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP’ 23, page 32–39. Association for Computing Machinery, 2023. |