Jaehyung **Seo**

KOREA UNIVERSITY · NLP AI LAB

145 Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea

📱 (+82) 10-7375-7552 | ✉ seojae777@korea.ac.kr | 🏠 j-seo.github.io/ | 🔘 J-Seo | 🔗 jaehyungseo-datascientist

Research Interests

I am currently a Ph.D candidate in the Department of Computer Science and Engineering at Korea University (Advisor: Heuiseok Lim). I am a co-founder and CTO of KU-NMT Group. I have published more than 50 papers in the field of Natural Language Processing, Computer Science, and applied for 4 patents.

- Natural Language Processing
- Artificial Intelligence
- Commonsense Reasoning
- Natural Language Generation
- Language Resource and Evaluation
- Language Modeling
- Decoding Strategy
- Neural Machine Translation
- Speech Recognition Post-Processor
- Data Management and Engineering
- Semantic Search and Information Retrieval

Education

Korea University   *Seoul, South Korea*

PH.D CANDIDATE IN COMPUTER SCIENCE AND ENGINEERING, GPA: 4.47 / 4.5)   *Sep. 2020 - Aug. 2022*

- **Advisor: Heuiseok Lim**
- **Natural Language Processing & Artificial Intelligence Lab**
- **Co-Founder and CTO of KU-NMT Group.**

Korea University   *Seoul, South Korea*

B.A. IN ENGLISH LANGUAGE AND LITERATURE & BUSINESS SCHOOL, GPA: 4.19 / 4.5   *Mar. 2014 - Exp. Aug. 2020*

- **Advisor: Donguk Kim**

Academic Services

Reviewer   *South Korea*

LANGUAGE RESOURCE AND EVALUATION & MULTILINGUAL   *October.2023*

- EMNLP 2023

Reviewer   *South Korea*

LANGUAGE RESOURCE AND EVALUATION & LANGUAGE MODELING   *January.2023*

- ACL 2023

Reviewer   *South Korea*

COMPUTATIONAL SOCIAL SCIENCE AND CULTURAL ANALYTICS   *December.2023*

- NAACL 2024

Reviewer   *South Korea*

GENERATION & LANGUAGE RESOURCE AND EVALUATION   *February.2024*

- ACL 2024

Selected Publications

KNOTICED: Augmentative and Alternative Communication Software for Language Developmental Disabilities

Sugyeong Eo, Jungwoo Lim, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim

LREC-COLING 2024 - The Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Leveraging Pre-existing Resources for Data-Efficient Counter-Narrative Generation in Korean
Seungyoon Lee, Chanjun Park, DaHyun Jung, Hyeonseok Moon, **Jaehyung Seo**, Sugyeong Eo, Heuiseok Lim
LREC-COLING 2024 - The Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Generative Interpretation: Toward Human-Like Evaluation for Educational Question-Answer Pair Generation
Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, **Jaehyung Seo**, Heuiseok Lim
EACL 2024 - Findings of the Association for Computational Linguistics

Hyper-BTS Dataset: Scalability and Enhanced Analysis of Back TranScription (BTS) for ASR Post-Processing
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Junyoung Son, Hyeonseok Moon, Sugyeong Eo, Chanhee Lee, Heuiseok Lim
EACL 2024 - Findings of the Association for Computational Linguistics

CReTIHC: Designing Causal Reasoning Tasks about Temporal Interventions and Hallucinated Confoundings
Changwoo Chun, SongEun Lee, **Jaehyung Seo**, Heui-Seok Lim
EMNLP 2023 - Findings of the Association for Computational Linguistics

CHEF in the Language Kitchen: A Generative Data Augmentation Leveraging Korean Morpheme Ingredients
Jaehyung Seo, Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, Heuiseok Lim
EMNLP 2023 - Proceedings of the Conference on Empirical Methods in Natural Language Processing

Informative Evidence-guided Prompt-based Fine-tuning for English-Korean Critical Error Detection
Dahyun Jung, Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heui-Seok Lim
IJCNLP-AACL 2023 - Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics

Knowledge Graph-Augmented Korean Generative Commonsense Reasoning
Dahyun Jung, **Jaehyung Seo**, Jaewook Lee, Chanjun Park, Heuiseok Lim
ICML-DMLR workshop (2023)

KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing
Seonmin Koo, Chanjun Park, Jinsung Kim, **Jaehyung Seo**, Sugyeong Eo, Hyeonseok Moon, Heui-Seok Lim
EMNLP 2023 - Proceedings of the Conference on Empirical Methods in Natural Language Processing

Uncovering the Risks and Drawbacks Associated with the Use of Synthetic Data for Grammatical Error Correction
Seonmin Koo, Chanjun Park, Seolhwa Lee, **Jaehyung Seo**, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
IEEE Access (2023). IEEE

PEEP-Talk: A Situational Dialogue-based Chatbot for English Education
Seungjun Lee, Yoonna Jang, Chanjun Park, Jungseob Lee, **Jaehyung Seo**, Hyeonseok Moon, Sugyeong Eo, Seounghoon Lee, Bernardo Yahya, Heui-Seok Lim
ACL 2023 - Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (*Volume 3: System Demonstrations*)

A Survey on Evaluation Metrics for Machine Translation
Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, Seonmin Koo, Heuiseok Lim
Mathematics 11.4 (2023) p. 1006. MDPI

Doubts on the reliability of parallel corpus filtering
Hyeonseok Moon, Chanjun Park, Seonmin Koo, Jungseob Lee, Seungjun Lee, **Jaehyung Seo**, Sugyeong Eo, Yoonna Jang, Hyunjoong Kim, Hyoung-gyu Lee
Expert Systems with Applications 233 (2023) p. 120962. Elsevier

Synthetic Alone: Exploring the Dark Side of Synthetic Data for Grammatical Error Correction
Chanjun Park, Seonmin Koo, Seolhwa Lee, **Jaehyung Seo**, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
ICML-DMLR workshop (2023)

QUAK: A Synthetic Quality Estimation Dataset for Korean-English Neural Machine Translation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Gyeongmin Kim, Jungseob Lee, Heui-Seok Lim
COLING 2022 - Proceedings of the 29th International Conference on Computational Linguistics

Word-level Quality Estimation for Korean-English Neural Machine Translation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim
IEEE Access 10 (2022) pp. 44964–44973. IEEE

A Dog Is Passing Over The Jet? A Text-Generation Dataset for Korean Commonsense Reasoning and Evaluation
Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, Heuiseok Lim
NAACL 2022 - Findings of the Association for Computational Linguistics, Seattle, United States

Dense-to-Question and Sparse-to-Answer: Hybrid Retriever System for Industrial Frequently Asked Questions
Jaehyung Seo, Taemin Lee, Hyeonseok Moon, Chanjun Park, Sugyeong Eo, Imatitikua D Aiyanyo, Kinam Park, Aram So, Sungmin Ahn, Jeongbae Park
Mathematics 10.8 (2022) p. 1335. MDPI

Plain Template Insertion: Korean-Prompt-Based Engineering for Few-Shot Learners
Jaehyung Seo, Hyeonseok Moon, Chanhee Lee, Sugyeong Eo, Chanjun Park, Jihoon Kim, Changwoo Chun, Heuiseok Lim
IEEE Access 10 (2022) pp. 107587–107597

PU-GEN: Enhancing generative commonsense reasoning for language models with human-centered knowledge
Jaehyung Seo, Dongsuk Oh, Sugyeong Eo, Chanjun Park, Kisu Yang, Hyeonseok Moon, Kinam Park, Heuiseok Lim
Knowledge-Based Systems (2022) p. 109861. Elsevier

Research on Utilization Strategy of User Engagements in Korean Fake News Detection
Myunghoon Kang, **Jaehyung Seo**, Chanjun Park, Heuiseok Lim
IEEE Access (2022) pp. 1–1

K-NCT: Korean Neural Grammatical Error Correction Gold-Standard Test Set Using Novel Error Type Classification Criteria
Seonmin Koo, Chanjun Park, **Jaehyung Seo**, Seungjun Lee, Hyeonseok Moon, Jungseob Lee, Heuiseok Lim
IEEE Access (2022). IEEE

BERTOEIC: Solving TOEIC Problems Using Simple and Efficient Data Augmentation Techniques with Pretrained Transformer Encoders
Jeongwoo Lee, Hyeonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, Heuiseok Lim
Applied Sciences 12.13 (2022) p. 6686. Multidisciplinary Digital Publishing Institute

Return on Advertising Spend Prediction with Task Decomposition-Based LSTM Model
Hyeonseok Moon, Taemin Lee, **Jaehyung Seo**, Chanjun Park, Sugyeong Eo, Imatitikua D Aiyanyo, Jeongbae Park, Aram So, Kyoungwha Ok, Kinam Park
Mathematics 10.10 (2022) p. 1637. MDPI

A Self-Supervised Automatic Post-Editing Data Generation Tool
Hyeonseok Moon, Chanjun Park, Sugyeong Eo, **Jaehyung Seo**, SeungJun Lee, Heuiseok Lim
ICML 2022 - DataPerf workshop (2022)

An Automatic Post Editing with Efficient and Simple Data Generation Method
Hyeonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, Heuiseok Lim
IEEE Access (2022). IEEE

Empirical Analysis of Noising Scheme based Synthetic Data Generation for Automatic Post-editing
Hyeonseok Moon, Chanjun Park, Seolhwa Lee, **Jaehyung Seo**, Jungseob Lee, Sugyeong Eo, Heuiseok Lim
LREC 2022 - Proceedings of the Language Resources and Evaluation Conference, Marseille, France

The ASR Post-Processor Performance Challenges of BackTranScription (BTS): Data-Centric and Model-Centric Approaches
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Chanhee Lee, Heuiseok Lim
Mathematics 10.19 (2022)

PicTalky: Augmentative and Alternative Communication Software for Language Developmental Disabilities
Chanjun Park, Yoonna Jang, Seolhwa Lee, **Jaehyung Seo**, Kisu Yang, Heuiseok Lim
IJCNLP-AACL 2022 - The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics

Priming Ancient Korean Neural Machine Translation
Chanjun Park, Seolhwa Lee, **Jaehyung Seo**, Hyeonseok Moon, Sugyeong Eo, Heuiseok Lim
LREC 2022 - Proceedings of the Language Resources and Evaluation Conference, Marseille, France

Empirical Analysis of Parallel Corpora and In-Depth Analysis Using LIWC
Chanjun Park, Midan Shim, Sugyeong Eo, Seolhwa Lee, **Jaehyung Seo**, Hyeonseok Moon, Heuiseok Lim
Applied Sciences 12.11 (2022) p. 5545. Multidisciplinary Digital Publishing Institute

A New Tool for Efficiently Generating Quality Estimation Datasets
Sugyeong Eo, Chanjun Park, **Jaehyung Seo**, Hyeonseok Moon, Heuiseok Lim
NeurIPS 2021 - Data-centric AI (*DCAI*) *workshop*

Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim
Applied Sciences 11.14 (2021) p. 6584. Multidisciplinary Digital Publishing Institute

Dealing with the Paradox of Quality Estimation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim
MT Summit - Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (*LoResMT2021*), Virtual

Automatic Knowledge Augmentation for Generative Commonsense Reasoning
Jaehyung Seo, Chanjun Park, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
NeurIPS 2021 - Data-centric AI (*DCAI*) *workshop*

Grounded Vocabulary for Image Retrieval Using a Modified Multi-Generator Generative Adversarial Network
Kuekyeng Kim, Chanjun Park, **Jaehyung Seo**, Heuiseok Lim
IEEE Access (2021). IEEE

An Empirical Study on Automatic Post Editing for Neural Machine Translation
Hyeonseok Moon, Chanjun Park, Sugyeong Eo, **Jaehyung Seo**, Heuiseok Lim
IEEE Access 9 (2021) pp. 123754–123763. IEEE

BTS: Back TranScription for Speech-to-Text Post-Processor using Text-to-Speech-to-Text
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Chanhee Lee, Hyeonseok Moon, Sugyeong Eo, Heuiseok Lim
ACL 2021 - Proceedings of the 8th Workshop on Asian Translation (*WAT2021*), Online

How should human translation coexist with NMT? Efficient tool for building high quality parallel corpus
Chanjun Park, Seolhwa Lee, Hyeonseok Moon, Sugyeong Eo, **Jaehyung Seo**, Heuiseok Lim
NeurIPS 2021 - Data-centric AI (*DCAI*) *workshop* (2021)

Honors & Awards

2023.10	**Best Paper Award, 2023**, The 35th Annual Conference on Human & Cognitive Language Technology	*South Korea*
2022.12	**NAVER PH.D Fellowship Award, 2022**, Scholarship Award	*South Korea*
2022.12	**1st place in WMT 2022 QE Task 3, 2022**, WMT 2022 - EMNLP 2022	*Abu Dhabi*
2022.10	**Best Paper Award, 2022**, The 34th Annual Conference on Human & Cognitive Language Technology	*South Korea*
2021.10	**Best Paper Award, 2021**, The 33rd Annual Conference on Human & Cognitive Language Technology	*South Korea*
2021.07	**4th Place in CommonGen 1.1 Leaderboard, 2021**, CommonGen 1.1 Leaderboard	*World Record*

Selected Research Experience

KOREA University *Seoul, South Korea*

RESEARCH IN REASONING AND GENERATION RESEARCH *June. 2021. - Present*

- CHEF in the Language Kitchen: A Generative Data Augmentation Leveraging Korean Morpheme Ingredients
- CReTIHC: Designing Causal Reasoning Tasks about Temporal Interventions and Hallucinated Confoundings
- Knowledge Graph-Augmented Korean Generative Commonsense Reasoning
- SRLev-BIH: An Evaluation Metric for Korean Generative Commonsense Reasoning
- Ko-ATOMIC: Korean Commonsense Knowledge Graph
- KommonGen: A Dataset for Korean Generative Commonsense Reasoning Evaluation
- Automatic Knowledge Augmentation for Generative Commonsense Reasoning
- Publication of research paper in EMNLP 2023: `https://aclanthology.org/2023.emnlp-main.367/`
- Publication of research paper in EMNLP-findings 2023: `https://aclanthology.org/2023.findings-emnlp.693/`
- Publication of research paper in NAACL 2022: `https://aclanthology.org/2022.findings-naacl.172/`
- Publication of research paper in Knowledge-Based Systems 2022: `https://doi.org/10.1016/j.knosys.2022.109861`
- Skills: Commonsense Reasoning, Pytorch, Huggingface

(Project) Next-Generation AI Technology Development - Hallucination *Seoul, South Korea*

RESEARCHER AT KOREA UNIV *Jan. 2024. - Dec. 2026*

- Supported by IITP
- Skills: Large language model, Hallucination detection/mitigation

(Project) Large Language Modeling for Special Technical Documents *Seoul, South Korea*

RESEARCHER AT KOREA UNIV *Dec. 2023. - June. 2024*

- Supported by GAON
- Skills: Large language modeling, RAG

(Project) Open Ko-LLM Leaderboard *Seoul, South Korea*

RESEARCHER AT KOREA UNIV *Sep. 2023. - Dec. 2023*

- Supported by Upstage
- Skills: Large language model, Commonsense Reasoning, Evaluation

(Project) UI/UX Artificial Intelligence Service Development *Seoul, South Korea*

RESEARCHER AT KOREA UNIV *June. 2023. - Feb. 2024.*

- Supported by Samsung Fire & Marine Insurance
- Skills: Natural language generation, Large language modeling, App application

(Project) R&D Voucher: AI Contact Center *Seoul, South Korea*

RESEARCHER AT KOREA UNIV *June. 2022. - May. 2023.*

- Supported by O2O
- Skills: Natural language processing, Pytorch, Huggingface

(Project) Industry-Academia-Government Collaboration R&D: AI Interviewing *Seoul, South Korea*

RESEARCHER AT KOREA UNIV *June. 2022. - May. 2023.*

- Supported by Withmind
- Skills: Natural language processing, Pytorch, Huggingface

KOREA University *Seoul, South Korea*

RESEARCH IN GRAMMATICAL ERROR CORRECTION *January. 2022. - Present*

- Uncovering the Risks and Drawbacks Associated with the Use of Synthetic Data for Grammatical Error Correction
- Synthetic Alone: Exploring the Dark Side of Synthetic Data for Grammatical Error Correction
- K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria
- Publication of research paper in IEEE Access (SCIE): `https://ieeexplore.ieee.org/abstract/document/9938990`
- Publication of research paper in IEEE Access (SCIE): `https://ieeexplore.ieee.org/abstract/document/10234394/`
- Publication of research paper in ICML 2023 Workshop: `https://arxiv.org/abs/2306.14377`
- Skills: Graph Neural Network, Pytorch, Fairseq

(Project) Data Augmentation with Language Generation Research

PROJECT MANAGER AT KOREA UNIV

Seoul, South Korea

May. 2022. - April. 2023.

- Supported by Samsung Research
- Skills: Data augmentation, Generation, Pytorch, Huggingface

(Project) Korean-Prompt-based Engineering for Few-shot Research

PROJECT MANAGER AT KOREA UNIV

Seoul, South Korea

March. 2021. - April. 2022.

- Plain Template Insertion: Korean-Prompt-based Engineering for Few-shot Learners
- Supported by Hyundai Motor
- Publication of research paper in IEEE Access (SCIE): https://ieeexplore.ieee.org/document/9913979
- Skills: Prompt learning, Few-shot learning, Pytorch, Huggingface

(Project) Return on Advertising Spend (ROAS) Prediction Research

RESEARCHER AT KOREA UNIV

Seoul, South Korea

December. 2021. - March. 2022.

- Return on Advertising Spend (ROAS) refers to the ratio of revenue generated by advertising projects to its expense.
- We present a method for predicting ROAS that does not require controlled experiments in data acquisition and validates its effectiveness through comparative experiments. Specifically, we propose a task deposition method that divides the end-to-end prediction task into the two-stage process; occurrence prediction and occurred ROAS regression.
- Supported by Data voucher
- Publication of research paper in Mathematics (SCIE): https://www.mdpi.com/2227-7390/10/10/1637
- Skills: ROAS, Regression, Pytorch

(Project) Hybrid Retriever System for Industrial Frequently Asked Questions Research

PROJECT MANAGER AT KOREA UNIV

Seoul, South Korea

December. 2021. - March. 2022.

- We propose a method of returning responses to customer inquiries using only data that can be easily obtained from companies based on the deep learning information retrieval module. In this work, we hybridize dense embedding and sparse embedding to make it robust to professional terms and suggest new functions to scale the results returned by the two modules and adjust the weight ratio.
- Supported by Data voucher
- Publication of research paper in Mathematics (SCIE): https://www.mdpi.com/2227-7390/10/8/1335
- Skills: Information Retrieval, Pytorch

KOREA University

RESEARCH IN FAKE NEWS DETECTION RESEARCH

Seoul, South Korea

January. 2022. - Present

- Research for Utilization Strategy of User Engagements in Korean Fake News Detection
- Publication of research paper in IEEE Access (SCIE): https://ieeexplore.ieee.org/document/9841541
- Skills: Graph Nerual Network, Pytorch, Fairseq

(Project) NAVER PAPAGO - Parallel Corpus Filtering and Mining

RESEARCHER AT KOREA UNIV

Seoul, South Korea

December. 2021. - June. 2022.

- Research for Static and Dynamic Parallel Corpus Filtering and Mining
- Human-inspired curriculum learning based NMT
- Supported by NAVER PAPAGO
- Skills: Neural Machine Translation, Parallel corpus filtering, Pytorch, Fairseq

KOREA University

RESEARCH IN DATA-CENTRIC NEURAL MACHINE TRANSLATION

Seoul, South Korea

July. 2021. - December. 2023.

- Doubts on the reliability of parallel corpus filtering
- Research for Data-centric Neural Machine Translation, Automatic Post Editing, Quality Estimation, Generative Commonsense Reasoning
- Develop Efficient tool for building high-quality parallel corpus and A New Tool for Efficiently Generating Quality Estimation Datasets and A Self-Supervised Automatic Post-Editing Data Generation Tool
- Publication of research paper in Expert Systems with Applications 2023: https://doi.org/10.1016/j.eswa.2023.120962
- Publication of 3 research paper in NeurIPS 2021 - Data-centric AI (DCAI) workshop: https://eval.how/dcai2021/index.html
- Publication of research paper in ICML 2022 - DataPerf workshop: https://sites.google.com/view/dataperf2022
- Skills: Neural Machine Translation, Web Application, Fairseq

KOREA University

RESEARCH IN IMAGE RETRIEVAL RESEARCH

Seoul, South Korea

June. 2021. - September. 2021.

- Grounded Vocabulary for Image Retrieval Using a Modified Multi-Generator Generative Adversarial Network
- Publication of research paper in IEEE Access (SCIE): https://ieeexplore.ieee.org/document/9585128
- Skills: Image Retrieval, Pytorch, Huggingface

(Project) Korean Multi-turn Response Chat-bot Research

PROJECT MANAGER AT KOREA UNIV

Seoul, South Korea

May. 2021. - January. 2022.

- Supported by Samsung SDS
- A study on Korean multi-turn response generation using generative and retrieval model: https://doi.org/10.15207/JKCS.2022.13.01.013
- Skills: Dialogue System, Natural language generation, Pytorch, Huggingface

KOREA University

Seoul, South Korea

RESEARCH IN SPEECH RECOGNITION POST-PROCESSOR RESEARCH

January. 2021. - Present

- Proposed BackTranScription(BTS). Back Transcription (BTS) has seen limited success without the role of phonetic transcription for building and creating vast datasets automatically.
- KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing
- Publication of research paper in EACL 2024
- Publication of research paper in EMNLP 2023: `https://aclanthology.org/2023.emnlp-main.292/`
- Publication of research paper in WAT 2021 (ACL Workshop): `https://aclanthology.org/2021.wat-1.10/`
- Publication of research paper in Mathematics: `https://www.mdpi.com/2227-7390/10/19/3618`
- Skills: Neural Machine Translation, Pytorch, Huggingface, Fairseq

KOREA University

Seoul, South Korea

RESEARCH IN AUTOMATIC POST EDITING RESEARCH

January. 2021. - Present

- APE research based on Cross Lingual Pretrained Language Model
- Adpater-based APE research
- Publication of research paper in IEEE Access (SCIE): `https://ieeexplore.ieee.org/document/9528385/`
- Research for Efficient and Simple Data Generation Method - Low Resource APE and Pseudo APE data generation
- Publication of research paper in IEEE Access (SCIE): `https://ieeexplore.ieee.org/document/9714400`
- Research for Synthetic Data Generation Using Noising Strategies for Automatic Post-editing
- Publication of research paper in LREC 2022: `http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.93.pdf`
- Skills: Neural Machine Translation, Pytorch, Automatic Post Editing, Huggingface, Fairseq

KOREA University

Seoul, South Korea

RESEARCH IN QUALITY ESTIMATION RESEARCH

January. 2021. - Present

- Large Scale PLM based Quality Estimation development and research
- Low Resource QE and Pseudo QE data generation
- Publication of research paper in Applied Science (SCIE): `https://www.mdpi.com/2076-3417/11/14/6584`
- Publication of research paper in MT Summit 2021 - LoresMT Workshop: `https://aclanthology.org/2021.mtsummit-loresmt.1/`
- Publication of research paper in IEEE Access: `https://ieeexplore.ieee.org/document/9761258`
- Publication of research paper in COLING 2022: `https://aclanthology.org/2022.coling-1.460.pdf`
- 1st place in WMT 2022 QE Task 3: KU X Upstage's submission for the WMT22 Quality Estimation: Critical Error Detection Shared Task
- Skills: Neural Machine Translation, Pytorch, Quality Estimation, OpenKiwi, Huggingface, Fairseq

KOREA University

Seoul, South Korea

RESEARCH IN AI FOR EDUCATION RESEARCH

September. 2020. - Present

- PEEP-Talk: Deep Learning-based English Education Platform for Personalized Foreign Language Learning
- AI student: A Machine Reading Comprehension System for the Korean College Scholastic Ability Test
- BERTOEIC: Solving TOEIC Problems Using Simple and Efficient Data Augmentation Techniques with Pretrained Transformer Encoders
- Publication of research paper in ACL 2023 (Demo-track): `https://aclanthology.org/2023.acl-demo.18/`
- Publication of research paper in Mathematics (SCIE): `https://www.mdpi.com/2227-7390/10/9/1486`
- Publication of research paper in Applied Sciences (SCIE): `https://www.mdpi.com/2076-3417/12/13/6686`
- Skills: Pytorch, Python, Google Cloud Platform, REST API, Nao Toolkit

KOREA University

Seoul, South Korea

RESEARCH IN AI FOR ACCESSIBILITY RESEARCH

September. 2020. - Present

- PicTalky: Design of augmentative alternative communication (AAC) software for children with developmental disabilities
- Publication of research paper in AACL 2022 Demo Track
- Skills: Pytorch, Python, Google Cloud Platform, REST API, Nao Toolkit

(Project) A Study on the Improvement of Understanding of Korean-based Generative Models

Seoul, South Korea

RESEARCHER AT KOREA UNIV

July. 2020. - March. 2021

- Supported by National Security Research Institute
- Skills: Language modeling, Natural language generation

Skills

Programming	Python
Machine Learning & Deep Learning	Pytorch, Pytorch lightning, Tensorflow
Natural Language Processing	Commonsense Reasoning, Machine Translation, Language Generation, Information Retrieval
GPU server Build & Management	Linux configuration and engineering, NAS, MLOps, GPU parallelization, Networking, RAID
OS	Linux, Ubuntu, Centos, MAC, Windows, Docker
Language Skills	Korean, English

Patents

2023.12	**DEVICE AND METHOD FOR KOREAN DATA AUGMENTATION BASED ON CONTRASTIVE LEARNING AND MORPHEME COMBINATION**, Patent Application (10-2023-0189496)	*South Korea*
2023.10	**APPARATUS AND METHOD FOR GENERATING LANGUAGE BASED ON COMMONSENSE REASONING**, Granted Patent (10-2593463)	*South Korea*
2023.03	**GENERATING METHOD OF MODEL MEASURING DOCUMENT SIMILARITY AND MEASURING METHOD FOR DOCUMENT SIMILARITY USING THE SAME**, Granted Patent (10-2507192)	*South Korea*
2022.10	**DEVICE AND METHOD FOR GENERATING KOREAN COMMONSENSE REASONING DATASET**, Patent Application (10-2022-0132373)	*South Korea*
2021.11	**APPARATUS OF OBTAINING RESPONSE BASED ON SEMANTIC SIMILARITY AND METHOD THEREOF**, Patent Application (10-2021-0156653)	*South Korea*

Invited Talk & Teaching Experience

Dongguk University *South Korea*

INVITED ADVISOR *Jun. 2023*

• Introduction to Natural Language Processing and Commonsense Reasoning

NongHyup Bank Co., Ltd *South Korea*

INVITED ADVISOR *Jul. 2020 - Sep. 2020*

• Teaching data science for new employees

Korea University *South Korea*

TEACHING ASSISTANT *Mar - Jun. 2023*

• (DFE615-00) Natural Language Processing for Artificial Intelligence

Korea University *South Korea*

TEACHING ASSISTANT *Sep - Dec. 2022*

• (DFE610) Natural Language Processing for Digital Finance Engineering
• Lecture materials: `https://github.com/nlpai-lab/nlp-bible-code`

Korea University *South Korea*

TEACHING ASSISTANT *Mar - Jun. 2022*

• (BDC101) Introduction to Natural Language Processing in Big Data
• Lecture materials: `https://github.com/J-Seo/BDC103_IR`

Korea University *South Korea*

RESEARCH ASSISTANT *Sep - Dec. 2021*

• (DFE610) Natural Language Processing for Digital Finance Engineering
• Lecture materials: `https://github.com/nlpai-lab/nlp-bible-code`

Korea University *South Korea*

TEACHING ASSISTANT *Mar - Jun. 2021*

• (BDC101) Introduction to Natural Language Processing in Big Data
• Lecture materials: `https://github.com/J-Seo/BDC103_IR`

All Publications

PREPRINTS

Language Chameleon: Transformation analysis between languages using Cross-lingual Post-training based on Pre-trained language models

Suhyune Son, Chanjun Park, Jungseob Lee, Midan Shim, Chanhee Lee, Yoonna Jang, **Jaehyung Seo**, Heuiseok Lim

arXiv preprint arXiv:2209.06422 (2022)

TOP CONFERENCE

KNOTICED: Augmentative and Alternative Communication Software for Language Developmental Disabilities

Sugyeong Eo, Jungwoo Lim, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim

LREC-COLING 2024 - The Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Leveraging Pre-existing Resources for Data-Efficient Counter-Narrative Generation in Korean

Seungyoon Lee, Chanjun Park, DaHyun Jung, Hyeonseok Moon, **Jaehyung Seo**, Sugyeong Eo, Heuiseok Lim

LREC-COLING 2024 - The Joint International Conference on Computational Linguistics, Language Resources and Evaluation

Generative Interpretation: Toward Human-Like Evaluation for Educational Question-Answer Pair Generation

Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, **Jaehyung Seo**, Heuiseok Lim

EACL 2024 - Findings of the Association for Computational Linguistics

Hyper-BTS Dataset: Scalability and Enhanced Analysis of Back TranScription (BTS) for ASR Post-Processing
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Junyoung Son, Hyeonseok Moon, Sugyeong Eo, Chanhee Lee, Heuiseok Lim
EACL 2024 - Findings of the Association for Computational Linguistics

CReTIHC: Designing Causal Reasoning Tasks about Temporal Interventions and Hallucinated Confoundings
Changwoo Chun, SongEun Lee, **Jaehyung Seo**, Heui-Seok Lim
EMNLP 2023 - Findings of the Association for Computational Linguistics

CHEF in the Language Kitchen: A Generative Data Augmentation Leveraging Korean Morpheme Ingredients
Jaehyung Seo, Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, Heuiseok Lim
EMNLP 2023 - Proceedings of the Conference on Empirical Methods in Natural Language Processing

Informative Evidence-guided Prompt-based Fine-tuning for English-Korean Critical Error Detection
Dahyun Jung, Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heui-Seok Lim
IJCNLP-AACL 2023 - Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics

KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing
Seonmin Koo, Chanjun Park, Jinsung Kim, **Jaehyung Seo**, Sugyeong Eo, Hyeonseok Moon, Heui-Seok Lim
EMNLP 2023 - Proceedings of the Conference on Empirical Methods in Natural Language Processing

PEEP-Talk: A Situational Dialogue-based Chatbot for English Education
Seungjun Lee, Yoonna Jang, Chanjun Park, Jungseob Lee, **Jaehyung Seo**, Hyeonseok Moon, Sugyeong Eo, Seounghoon Lee, Bernardo Yahya, Heui-Seok Lim
ACL 2023 - Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)

QUAK: A Synthetic Quality Estimation Dataset for Korean-English Neural Machine Translation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Gyeongmin Kim, Jungseob Lee, Heui-Seok Lim
COLING 2022 - Proceedings of the 29th International Conference on Computational Linguistics

A Dog Is Passing Over The Jet? A Text-Generation Dataset for Korean Commonsense Reasoning and Evaluation
Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, Heuiseok Lim
NAACL 2022 - Findings of the Association for Computational Linguistics, Seattle, United States

PicTalky: Augmentative and Alternative Communication Software for Language Developmental Disabilities
Chanjun Park, Yoonna Jang, Seolhwa Lee, **Jaehyung Seo**, Kisu Yang, Heuiseok Lim
IJCNLP-AACL 2022 - The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics

Priming Ancient Korean Neural Machine Translation
Chanjun Park, Seolhwa Lee, **Jaehyung Seo**, Hyeonseok Moon, Sugyeong Eo, Heuiseok Lim
LREC 2022 - Proceedings of the Language Resources and Evaluation Conference, Marseille, France

INTERNATIONAL JOURNAL

Uncovering the Risks and Drawbacks Associated with the Use of Synthetic Data for Grammatical Error Correction
Seonmin Koo, Chanjun Park, Seolhwa Lee, **Jaehyung Seo**, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
IEEE Access (2023). IEEE

A Survey on Evaluation Metrics for Machine Translation
Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, Seonmin Koo, Heuiseok Lim
Mathematics 11.4 (2023) p. 1006. MDPI

Doubts on the reliability of parallel corpus filtering
Hyeonseok Moon, Chanjun Park, Seonmin Koo, Jungseob Lee, Seungjun Lee, **Jaehyung Seo**, Sugyeong Eo, Yoonna Jang, Hyunjoong Kim, Hyoung-gyu Lee
Expert Systems with Applications 233 (2023) p. 120962. Elsevier

Word-level Quality Estimation for Korean-English Neural Machine Translation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim
IEEE Access 10 (2022) pp. 44964–44973. IEEE

Dense-to-Question and Sparse-to-Answer: Hybrid Retriever System for Industrial Frequently Asked Questions
Jaehyung Seo, Taemin Lee, Hyeonseok Moon, Chanjun Park, Sugyeong Eo, Imatitikua D Aiyanyo, Kinam Park, Aram So, Sungmin Ahn, Jeongbae Park
Mathematics 10.8 (2022) p. 1335. MDPI

Plain Template Insertion: Korean-Prompt-Based Engineering for Few-Shot Learners
Jaehyung Seo, Hyeonseok Moon, Chanhee Lee, Sugyeong Eo, Chanjun Park, Jihoon Kim, Changwoo Chun, Heuiseok Lim
IEEE Access 10 (2022) pp. 107587–107597

PU-GEN: Enhancing generative commonsense reasoning for language models with human-centered knowledge
Jaehyung Seo, Dongsuk Oh, Sugyeong Eo, Chanjun Park, Kisu Yang, Hyeonseok Moon, Kinam Park, Heuiseok Lim
Knowledge-Based Systems (2022) p. 109861. Elsevier

Research on Utilization Strategy of User Engagements in Korean Fake News Detection
Myunghoon Kang, **Jaehyung Seo**, Chanjun Park, Heuiseok Lim
IEEE Access (2022) pp. 1–1

BERTOEIC: Solving TOEIC Problems Using Simple and Efficient Data Augmentation Techniques with Pretrained Transformer Encoders
Jeongwoo Lee, Hyeonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, Heuiseok Lim
Applied Sciences 12.13 (2022) p. 6686. Multidisciplinary Digital Publishing Institute

Return on Advertising Spend Prediction with Task Decomposition-Based LSTM Model
Hyeonseok Moon, Taemin Lee, **Jaehyung Seo**, Chanjun Park, Sugyeong Eo, Imatitikua D Aiyanyo, Jeongbae Park, Aram So, Kyoungwha Ok, Kinam Park
Mathematics 10.10 (2022) p. 1637. MDPI

An Automatic Post Editing with Efficient and Simple Data Generation Method
Hyeonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, Heuiseok Lim
IEEE Access (2022). IEEE

The ASR Post-Processor Performance Challenges of BackTranScription (BTS): Data-Centric and Model-Centric Approaches
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Chanhee Lee, Heuiseok Lim
Mathematics 10.19 (2022)

Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim
Applied Sciences 11.14 (2021) p. 6584. Multidisciplinary Digital Publishing Institute

Grounded Vocabulary for Image Retrieval Using a Modified Multi-Generator Generative Adversarial Network
Kuekyeng Kim, Chanjun Park, **Jaehyung Seo**, Heuiseok Lim
IEEE Access (2021). IEEE

An Empirical Study on Automatic Post Editing for Neural Machine Translation
Hyeonseok Moon, Chanjun Park, Sugyeong Eo, **Jaehyung Seo**, Heuiseok Lim
IEEE Access 9 (2021) pp. 123754–123763. IEEE

Domestic Journal

Policy-based performance comparison study of Real-time Simultaneous Translation
Jungseob Lee, Hyeonseok Moon, Chanjun Park **Jaehyung Seo**, Sugyeong Eo, Seungjun Lee, Seonmin Koo, Heuiseok Lim
Journal of the Korea Convergence Society 13.3 (2022) pp. 43–54. Korea Convergence Society

Error Analysis of Recent Conversational Agent-based Commercialization Education Platform
Seungjun Lee, Chanjun Park, **Jaehyung Seo**, Heuiseok Lim
Journal of the Korea Convergence Society 13.3 (2022) pp. 11–22. Korea Convergence Society

Study on Zero-shot based Quality Estimation
Sugyeong Eo, Chanjun Park, **Jaehyung Seo**, Hyeonseok Moon, Heuiseok Lim
Journal of the Korea Convergence Society 12.11 (2021) pp. 35–43. Korea Convergence Society

Research on Recent Quality Estimation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, HeuiSeok Lim
Journal of the Korea Convergence Society 12.7 (2021) pp. 37–44. Korea Convergence Society

Study on Decoding Strategies in Neural Machine Translation
Jaehyung Seo, Chanjun Park, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
Journal of the Korea Convergence Society 12.11 (2021) pp. 69–80. Korea Convergence Society

Recent Automatic Post Editing Research
Hyeonseok Moon, Chanjun Park, Sugyeong Eo, **Jaehyung Seo**, HeuiSeok Lim
Journal of Digital Convergence 19.7 (2021) pp. 199–208. Journal of Digital Convergence

The Verification of the Transfer Learning-based Automatic Post Editing Model
Hyeonseok Moon, Chanjun Park, Sugyeong Eo, **Jaehyung Seo**, Heuiseok Lim
Journal of the Korea Convergence Society 12.10 (2021) pp. 27–35. Korea Convergence Society

A Study on Verification of Back TranScription(BTS)-based Data Construction
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Hyeonseok Moon, Sugyeong Eo, Heuiseok Lim
Journal of the Korea Convergence Society 12.11 (2021) pp. 109–117. Korea Convergence Society

Domestic Conference

Word-level Korean-English Quality Estimation
Sugyeong Eo, Chanjun Park, **Jaehyung Seo**, Heonseok Moon, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

KommonGen: A Dataset for Korean Generative Commonsense Reasoning Evaluation (Best Paper Award)
Jaehyung Seo, Chanjun Park, Heonseok Moon, Sugyeong Eo, Myunghoon Kang, Seounghoon Lee, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

Persona-based Korean Conversational Model
Yoonna Jang, Jungwoo Lim, Yuna Hur, Kisu Yang, Chanjun Park, **Jaehyung Seo**, Seungjun Lee, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

Proposal of Automated Essay Scoring Method based on Deep-Learning
Yujin Kim, Chanjun Park, Seolhwa Lee, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

Model Training and Data Augmentation Schemes For the High-level Machine Reading Comprehension
Jeongwoo Lee, Hyeonseok Moon, Chanjun Park, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

Comparative study of Korean speech recognition based on SpecAugment and Kaldi
Seounghoon Lee, Chanjun Park, **Jaehyung Seo**, Gyeongmin Kim, HeuiSeok Lim

The 33rd Annual Conference on Human Cognitive Language Technology

Verification of the Domain Specialized Automatic Post Editing Model
Heonseok Moon, Chanjun Park, **Jaehyung Seo**, Sugyeong Eo, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

BackTranScription (BTS)-based Jeju Automatic Speech Recognition Post-processor Research
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Heonseok Moon, Sugyeong Eo, Yoonna Jang, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

Categorization and Analysis of Error Types in the Korean Speech Recognition System (Best Paper Award)
Junyoung Son, Chanjun Park, **Jaehyung Seo**, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

Attention Patterns and Semantics of Korean Language Models
Kisu Yang, Yoonna Jang, Jingwoo Lim, Chanjun Park, Hwanseok Jang, HeuiSeok Lim
The 33rd Annual Conference on Human Cognitive Language Technology

MMA: Multi-modal Message Aggregation for Korean VQA
Sungjin Park, Chanjun Park, **Jaehyung Seo**, HeuiSeok Lim
The 32st Annual Conference on Human Cognitive Language Technology

WORKSHOP

Knowledge Graph-Augmented Korean Generative Commonsense Reasoning
Dahyun Jung, **Jaehyung Seo**, Jaewook Lee, Chanjun Park, Heuiseok Lim
ICML-DMLR workshop (2023)

Synthetic Alone: Exploring the Dark Side of Synthetic Data for Grammatical Error Correction
Chanjun Park, Seonmin Koo, Seolhwa Lee, **Jaehyung Seo**, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
ICML-DMLR workshop (2023)

Focus on FoCus: Is FoCus focused on Context, Knowledge and Persona?
SeungYoon Lee, Jungseob Lee, Chanjun Park, Sugyeong Eo, Hyeonseok Moon, **Jaehyung Seo**, Jeongbae Park, Heui-Seok Lim
COLING 2022 - Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge

A Self-Supervised Automatic Post-Editing Data Generation Tool
Hyeonseok Moon, Chanjun Park, Sugyeong Eo, **Jaehyung Seo**, SeungJun Lee, Heuiseok Lim
ICML 2022 - DataPerf workshop (2022)

A New Tool for Efficiently Generating Quality Estimation Datasets
Sugyeong Eo, Chanjun Park, **Jaehyung Seo**, Hyeonseok Moon, Heuiseok Lim
NeurIPS 2021 - Data-centric AI (*DCAI*) *workshop*

Dealing with the Paradox of Quality Estimation
Sugyeong Eo, Chanjun Park, Hyeonseok Moon, **Jaehyung Seo**, Heuiseok Lim
MT Summit - Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (*LoResMT2021*), Virtual

Automatic Knowledge Augmentation for Generative Commonsense Reasoning
Jaehyung Seo, Chanjun Park, Sugyeong Eo, Hyeonseok Moon, Heuiseok Lim
NeurIPS 2021 - Data-centric AI (*DCAI*) *workshop*

BTS: Back TranScription for Speech-to-Text Post-Processor using Text-to-Speech-to-Text
Chanjun Park, **Jaehyung Seo**, Seolhwa Lee, Chanhee Lee, Hyeonseok Moon, Sugyeong Eo, Heuiseok Lim
ACL 2021 - Proceedings of the 8th Workshop on Asian Translation (*WAT2021*), Online

How should human translation coexist with NMT? Efficient tool for building high quality parallel corpus
Chanjun Park, Seolhwa Lee, Hyeonseok Moon, Sugyeong Eo, **Jaehyung Seo**, Heuiseok Lim
NeurIPS 2021 - Data-centric AI (*DCAI*) *workshop* (2021)