

First International Workshop on Knowledge-Enhanced Machine Translation

Proceedings of the Workshop

June 27, 2024

The KEMT organizers gratefully acknowledge the support from the following sponsors.

Sponsors of the Workshop





The papers published in this proceedings are —unless indicated otherwise— covered by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC-BY-NCND 4.0). You may copy, distribute, and transmit the work, provided that you attribute it (authorship, proceedings, publisher) in the manner specified by the author(s) or licensor(s), and that you do not use it for commercial purposes. The full text of the licence may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en

©2024 The authors

ISBN 978-1-0686907-4-7

Introduction

This volume contains the proceedings of the First International Workshop on Knowledge-Enhanced Machine Translation (KEMT 2024), hosted by the 25th Annual Conference of the European Association for Machine Translation (EAMT 2024). KEMT 2024 focuses on all aspects of the integration of additional knowledge into machine translation, including translation memories, terminology, linguistic information, translation quality indicators ...

The workshop welcomed submissions of either research papers or extended abstracts/industry reports. Three research papers and two extended abstracts were received. Each submission was reviewed by three independent members of the program committee, and the final decision was made by the program chairs. The five submissions were accepted. The three research papers are to be presented orally, and the two extended abstracts will be presented as posters.

The accepted papers cover a diverse range of ways to add external information to machine translation systems: terminology, lexicons, fuzzy matches, hypernyms, and even language grammars.

In addition to the research papers and abstracts, we are honored to have two invited speakers from the industry: Ricardo Rei (Unbabel), with a keynote entitled "TowerLLM: Improving Translation Quality through Prompting with Terminology and Translation Guidelines"; and Tom Vanallemeersch (CrossLang) with the keynote "To Customize Is to Know: Leveraging In-house Knowledge for Multilingual Document Flows."

Finally, the program includes a panel discussion where participants can share their thoughts about many aspects of the integration of external information into machine translation: the needs of industry and trending research topics, how large language models are changing the landscape, etc.

We sincerely thank all the people and institutions that contributed to the success of the workshop: the authors of the submitted papers for their interest in the topic, the program committee members for their valuable feedback and insightful comments; and the EAMT organizers for their support. We also thank our sponsors, Ghent University and the Language and Translation Technology Team (LT3), for their generous contributions.

Arda Tezcan, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis Workshop Organizers

Organizing Committee

Members of the Organizing Committee and Program Chairs

Arda Tezcan, Universiteit Gent, Belgium Víctor M. Sánchez-Cartagena, Universitat d'Alacant, Spain Miquel Esplà-Gomis, Universitat d'Alacant, Spain

Program Committee

Program Committee

Frédéric Blain, Tilburg University, the Netherlands Josep Crego, Systran, France Miquel Esplà-Gomis, Universitat d'Alacant, Spain Yasmin Moslem, Dublin City University, Ireland Juan Antonio Pérez-Ortiz, Universitat d'Alacant, Spain Víctor M. Sánchez-Cartagena, Universitat d'Alacant, Spain Felipe Sánchez-Martínez, Universitat d'Alacant, Spain Arda Tezcan, Universiteit Gent, Belgium Torregrosa Torregrosa, World Intellectual Property Organization, Switzerland Antonio Toral, University of Groningen, the Netherlands Tom Vanallemeersch, CrossLang, Belgium Vincent Vandeghinste, Instituut voor de Nederlandse Taal, the Netherlands Bram Vanroy, Katholieke Universiteit Leuven, Belgium François Yvon, CNRS and Sorbonne-Université, France

Invited Speakers

Ricardo Rei, Unbabel (Lisbon, Portugal) Tom Vanallemeersch, CrossLang (Gent, Belgium)

Keynote Talk To Customize Is to Know: Leveraging In-house Knowledge for Multilingual Document Flows

Tom Vanallemeersch CrossLang (Gent, Belgium) 2024-06-27 09:05:00 – Room: KEMT Workshop room

Abstract: While the number of commercial and open-source multilingual NLP models steadily keeps growing, such generic models do not necessarily meet users' unique demands in full. This is especially true for companies and public administrations with highly specialized document flows. To optimize the use of multilingual tools, these organizations should be aware of the value of their in-house knowledge. This knowledge is not only embedded in multilingual assets like translation memories, documents in various languages and formats, or glossaries, but also the in-house expertise on document functionality and critical textual elements like terms and named entities.

Bio: Tom Vanallemeersch is Language AI Adviser at CrossLang, where he contributes to the customisation and deployment of multilingual NLP systems and coordinates the company's participation in publicly funded projects. Besides various positions in industry, including work at Systran, his career spans academia (PhD in computational linguistics at the University of Leuven) and consultancy for the European Commission (DG Translation's MT team). In his spare time, his membership of a chamber choir allows him to conduct multilingual experiments of a wholly different kind.

Keynote Talk

TowerLLM: Improving Translation Quality through Prompting with Terminology and Translation Guidelines

Ricardo Rei Unbabel (Lisbon, Portugal)

2024-06-27 11:00:00 – Room: KEMT Workshop room

Abstract: TowerLLM revolutionizes machine translation by tailoring large language models (LLMs) to diverse translation tasks. By continued pretraining on mixed data and fine-tuning with task-specific instructions, TowerLLM surpasses open alternatives and rivals closed LL-Ms. This approach ensures proficiency across translation workflows, enhancing quality and efficiency. TowerLLM's impact extends beyond technical advancements, envisioning a future where specialized LLMs seamlessly integrate into translation pipelines, augmenting human capabilities. With the release of Tower models, specialized datasets, and evaluation frameworks, TowerLLM democratizes access to specialized resources, fostering collaboration and driving transformative advancements in machine translation.

Bio: Ricardo Rei is a senior research scientist at Unbabel, specializing in machine translation and natural language processing. He is set to complete his Ph.D. in April, which has been a collaborative effort between Unbabel, INESC-ID/Tecnico, and CMU University. His doctoral research has been centered on machine translation evaluation, and he is the main developer behind the COMET evaluation framework, which has become the industry standard metric for assessing machine translation quality. With a keen interest in advancing the capabilities of multilingual large language models (LLMs), he has been at the forefront of research and development in this domain. When not immersed in research, Ricardo enjoys maintaining an active lifestyle, often found at the gym or riding the waves while surfing—a passion he has pursued since the age of nine.

Table of Contents

Incorporating Hypernym Features for Improving Low-resource Neural Machine Translation Abhisek Chakrabarty, Haiyue Song, Raj Dabre, Hideki Tanaka and Masao Utiyama1
<i>Exploring Inline Lexicon Injection for Cross-Domain Transfer in Neural Machine Translation</i> Jesujoba O. Alabi and Rachel Bawden
Adding soft terminology constraints to pre-trained generic MT models by means of continued training
Tommi Nieminen
Leveraging Synthetic Monolingual Data for Fuzzy-Match Augmentation in Neural Machine Translation: A Preliminary Study
Thomas Moerman and Arda Tezcan34
Can True Zero-shot Methods with Large Language Models be Adopted for Sign Language
Machine Translation?
Euan McGill and Horacio Saggion40