Karlsruhe Institute of Technology

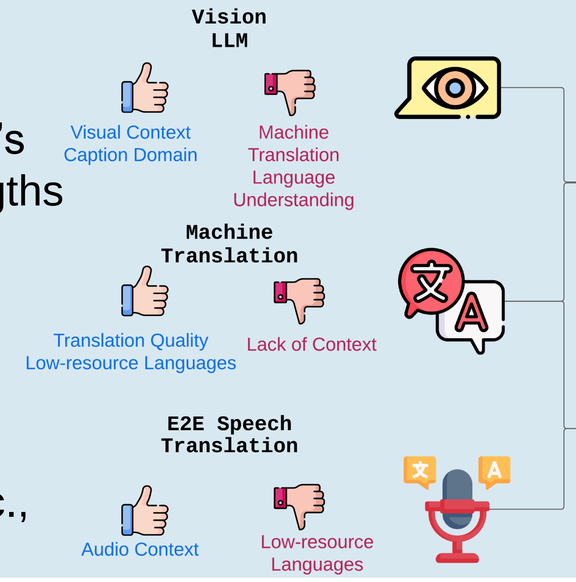
Artificial Intelligence for Language Technologies (AI4LT) Group

Plug, Play, and Fuse: Zero-Shot Joint Decoding via Word-Level Re-ranking Across Diverse Vocabularies

Sai Koneru, Matthias Huck, Miriam Exel, Jan Niehues
{sai.koneru,jan.niehues}@kit.edu, {matthias.huck, miriam.exel}@sap.com

Motivation

- Increase in open-source LLM's
- Equipped with different strengths
  - Multimodal, Chat bot, Legal etc.,
- Fuse models for combining strengths
  - Integrate contextual information, robustness, etc.,

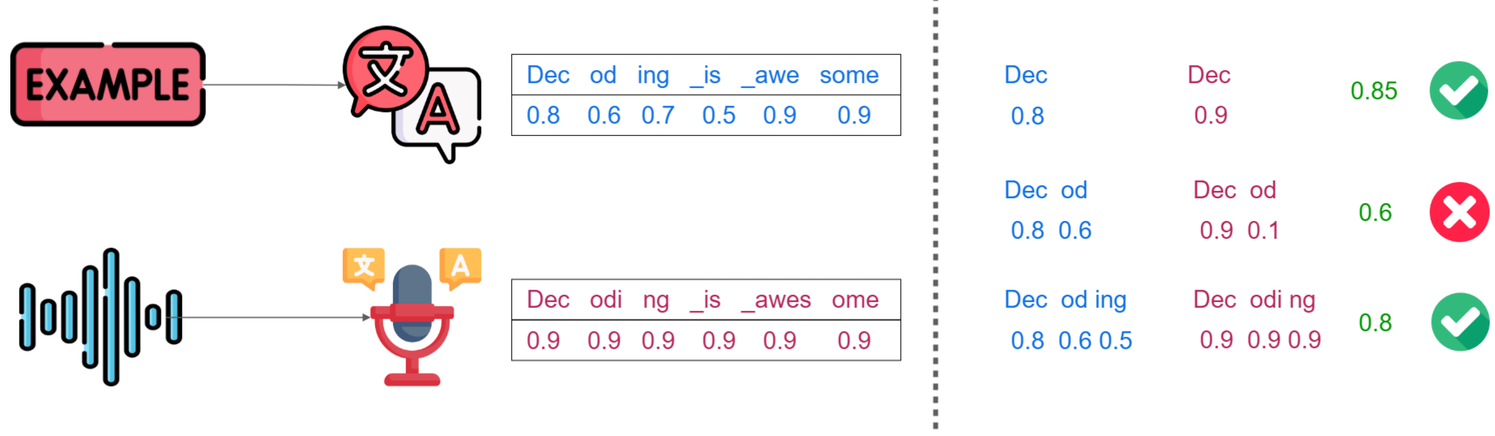


Challenges

- Handle diverse vocabularies
- Zero-shot Joint Decoding
  - No Training
  - No Data
- Traditional re-ranking is insufficient
  - Early influence is crucial
  - Partial re-ranking - incorrect estimates

Approach

- Word-level Online re-ranking
- Get top candidates from Generator
- Merge score for each candidate
- Use ranker model to predict end of word
  - Yes – rank full candidate
  - No – rank preceding words
- Length re-normalization for accurate scores



Experimental Setup

- Unimodal Translation
- Multimodal Translation

Test Set	Examples	Language	Phenomenon
WMT 2022	2037	En → De	Translation
MuST-SHE	315 (1108)	En → Fr	Gender Disambiguation Translation
CoMMuTE	300	En → De	Image Disambiguation Translation

Example: Speech-aware Translation via joint decoding

Source: As an undergraduate student, I **fell** in love with one of them... **Mrs. Ples**
Reference: Quand j ' étais étudiante en licence, je suis tombée amoureuse de l'un d'eux … Mme Ples.
Seamless Bal: En tant qu'étudiante de premier cycle, je suis tombée amoureuse d'une d'elles, Mme Platt.
Madlad: Comme un étudiant de premier cycle, je suis tombé amoureux de l'un d'eux... Mme Ples.
Joint Decoding: En tant qu'étudiante de premier cycle, je suis tombée amoureuse de l'une d'entre elles, Mme Ples.

Results: Unimodal Translation

- Fusing models improves translation quality
- Online re-ranking provides accurate probabilities
- Integrating with QE pushes the quality further

Generator	Ranker	Online	XCOMET-XXL
GPT-4	N/A	N/A	97.56
Madlad	N/A	N/A	96.77
Alma-R	N/A	N/A	97.48
Madlad	Alma-R	×	97.12
Madlad, Alma-R	Alma-R, Madlad	×	97.39
Madlad	Alma-R	✓	**97.68**
QE re-ranking			
Madlad	N/A	N/A	97.25
Madlad	Alma-R	✓	**97.91**

Results: Multimodal Translation

- Madlad and Seamless complement each other
- Joint Decoding achieves balance between translation quality and gender disambiguation
- No improvements over traditional re-ranking in image-aware translation

Gen.	Rank.	Onl.	1F (%)	1M (%)	COMET-22
Seamless Bal	N/A	N/A	**50**	66	80.48
Madlad	N/A	N/A	26	**90**	83.52
Madlad	Seamless Bal	×	29	**90**	83.66
Seamless Bal	Madlad	×	41	78	81.31
Madlad	Seamless Bal	✓	34	87	**83.78**

Limitations

- Character-level languages
- Latency and Memory

- Current approach combines strengths but also weaknesses
  - How to dynamically decide when to trust which model?
  - Contrastive decoding?