Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting

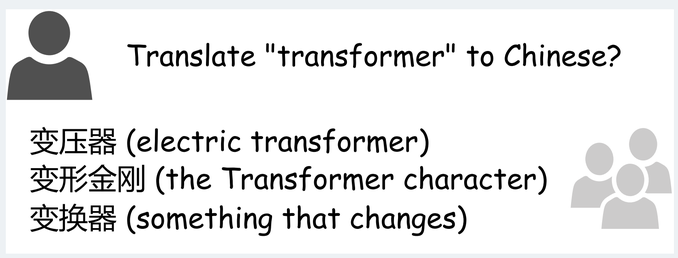
Nikolay Bogoychev* Pinzhen Chen*

School of Informatics, University of Edinburgh

{n.bogoych,pinzhen.chen}@ed.ac.uk *Equal contribution

Overview

- **UEdin@WMT23 Terminology shared task**
  - Terminology hints can help to disambiguate when translating polysemantic words with limited context.
  - Translate-then-refine, minimal manual efforts.



  - **Three language pairs**:
    German→English, Chinese→English and English→Czech
  - **Two baseline translation methods**:
    Terminology-aware translation & LLM translation.
  - **Two refinement methods**:
    Negatively constrained decoding & LLM refinement.

Translate

- **Terminology-aware translation (TAT)**
  - Compute word alignments through IBM methods.
  - Augment with 7% random pseudo-terminology alignments.
    Where is the airport? ↔ Wo ist der Flughafen?
    Where is the airport __*target*__ *Flughafen* __*done*__? ↔ Wo ist der Flughafen?
  - Train Transformer-big.
  - Translate by inserting the terminology hints into the source.
- **Ask GPT nicely to translate**
  - Use gpt-3.5-turbo API to translate without terminology constraints.

Refine

- **Negative constrained decoding (NCD)**
  - Check if all terminology words are accounted for.
  - If not, perform word alignments to identify missed terms.
  - Re-translate but blacklist the wrongly produced target words.
- **Ask GPT nicely to refine**
  - Use chatGPT API to refine the output with terminology.

Query	Prompt template
Translation	`Source: ${source}` `Please give me a translation in ${lang} without any explanation.`

Refinement	`Source: ${source}`
	`Bad Translation: ${translation}`
	`Please give me a better ${lang} translation without any explanation.`
	``` ``${srcword$_0$}'' should be translated as ``${trgword$_0$}'';
	``${srcword$_1$}'' should be translated as ``${trgword$_1$}'';
	...
	``${srcword$_k$}'' should be translated as ``${trgword$_k$}''. *(with k >= 0)* ```

Table: Prompts used for ChatGPT translation and terminology refinement.

Internal Evaluation

- **We computed terminology recall and quality internally**
  - Terminology-aware translation is stronger for German→English.
  - LLM refinement is effective, and it is stronger for Chinese→English and English→Czech.

Mode	Model	Refine	de→en Recall	de→en $COMET_{QE}$	zh→en Recall	zh→en $COMET_{QE}$	en→cs Recall	en→cs $COMET_{QE}$
terminology constraints	TAT	-	**82.30**	.0797	49.98	-.0896	73.75	.0601
	TAT	NCD	82.01	.0775	50.42	-.0903	73.26	.0588
	TAT	LLM	64.35	.1197	**83.06**	.0185	76.00	.0866
	LLM	-	41.86	.1244	46.63	.0191	48.14	.0913
	LLM	LLM	70.48	.1180	81.01	.0201	**78.94**	.0882
no constraint†	TAT	-	39.82	.1085	13.64	-.1163	48.11	.0712
	TAT	LLM	39.59	.1251	42.76	.0203	47.31	**.0955**
	LLM	-	41.86	.1244	46.63	.0191	48.14	.0913
	LLM	LLM	39.65	**.1258**	46.72	**.0228**	46.22	.0943
random constraints	TAT	-	**76.17**	.0716	81.55	-.1105	57.10	.0502
	TAT	NCD	75.79	.0698	**82.03**	-.1123	56.42	.0465
	TAT	LLM	61.46	.1206	63.17	.0175	70.97	.0875
	LLM	-	38.70	.1244	52.49	.0191	39.34	.0913
	LLM	LLM	66.74	.1188	67.10	.0196	**73.37**	.0867
no constraint‡	TAT	-	35.60	.1085	36.18	-.1163	37.35	.0712
	TAT	LLM	37.58	.1251	49.48	.0203	39.03	**.0955**
	LLM	-	38.70	.1244	52.49	.0191	39.34	.0913
	LLM	LLM	37.62	**.1258**	49.00	**.0228**	38.42	.0943

†Recall computed against terminology constraints.
‡Recall computed against random constraints.

Table: Terminology recall and translation quality measured by $COMET_{QE}$ of our systems on the *blind test* set.
TAT: terminology-aware translation; NCD: negatively constrained decoding; LLM: large language model.

How did we do?

- **Official evaluation results**
  - Our systems achieved the best overall quality.

System	$Comet^{DA}_{22}$ De←En	$Comet^{DA}_{22}$ En←Cs	$Comet^{DA}_{22}$ Zh←En
$UEDIN_{LLM}$	**0.813**	**0.869**	**0.757**
$UEDIN_{TAT}$	0.809	**0.868**	**0.757**
OPUS-CAT	0.790	**0.869**	0.521
AdaptTerm	0.801	0.841	0.688
$UEDIN_{NCD}$	0.792	0.835	0.650
LinguaCustodia	0.735	0.834	0.609
$VARCO\text{-}MT_{TSSNMT}$			0.755
BJTU-LB			0.751
$VARCO\text{-}MT_{ForceGen}$			0.715

Table: Translation quality

  - We occasionally miss some terminology constraints as our methods cannot enforce their appearance.

System	Terminology success rate De←En	Terminology success rate En←Cs	Terminology success rate Zh←En
AdaptTerm	**0.587**	0.613	0.758
Lingua Custodia	**0.622**	**0.662**	0.747
OPUS-CAT	0.443	0.557	0.124
$UEDIN_{LLM}$	0.560	0.629	0.753
$UEDIN_{TAT}$	0.539	0.626	0.739
$UEDIN_{NCD}$	0.587	0.562	0.536
$VARCO\text{-}MT_{TSSNMT}$			0.779
$VARCO\text{-}MT_{ForceGen}$			**0.800**
BJTU-LB			0.749

Table: Terminology success rate

THE UNIVERSITY of EDINBURGH informatics
High Performance Language Technologies
UTTER Unified Transcription and Translation for Extended Reality
UK Research and Innovation