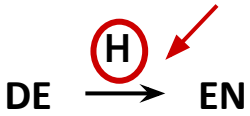
Mitigating Translationese with GPT-4: Strategies and Performance

Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl,
Cristina España-Bonet, and Josef van Genabith

Department of Language Science and Technology, Saarland University
German Research Center for Artificial Intelligence (DFKI)

What is Translationese?

Set of features unique to translated texts, found even in high-quality translations (Gellerstam, 1986; Toury, 1980; Baker, 1993)



Human translators' translation direction

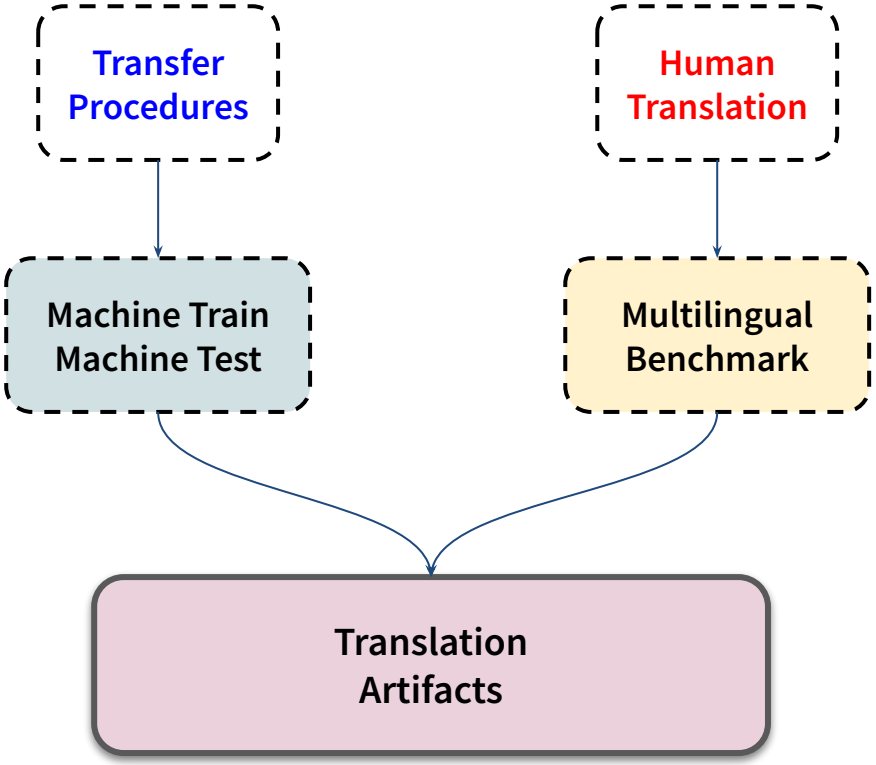
Original DE	**In der Innenstadt explodierte eine Autobombe**
Translated EN	**In the inner city there exploded a car bomb**

More natural translation: **A car bomb exploded downtown**

Lexical and structural "shining-through" [Teich, 2003]

Example inspired from "Translation, Translation Data, and Evaluation" (Graham Neubig)

Need of Mitigating Translationese



Such artefacts have a notable impact in the performance of cross-lingual applications

The Effect of Translationese in Machine Translation Test Sets
(Zhang and Toral, 2019 WMT)

BLEU might be Guilty but References are not Innocent
(Freitag et al., 2020 EMNLP)

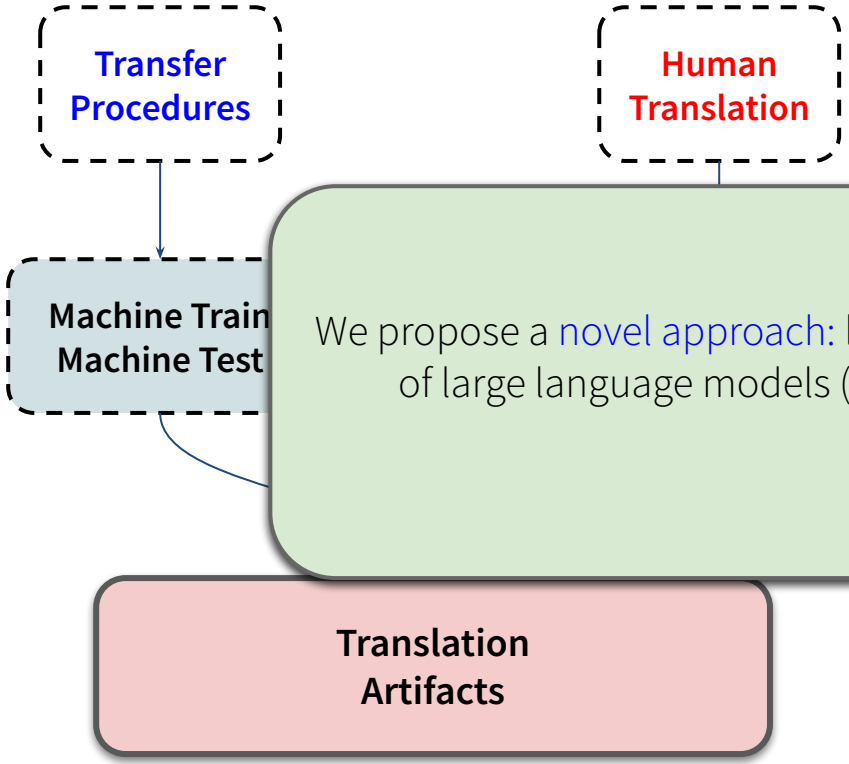
Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation
(Toral et al., 2018 WMT)

Statistical Power and Translationese in Machine Translation Evaluation
(Graham et al., 2020 EMNLP)

Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation
(Läubli et al., 2018 EMNLP)

Findings of the 2019 Conference on Machine Translation (WMT19)
(Freitag et al., 2019 WMT)

Need of Translationese Debiasing



Such artefacts have a notable impact in the performance of cross-lingual applications

ural Machine

t-level Evaluation

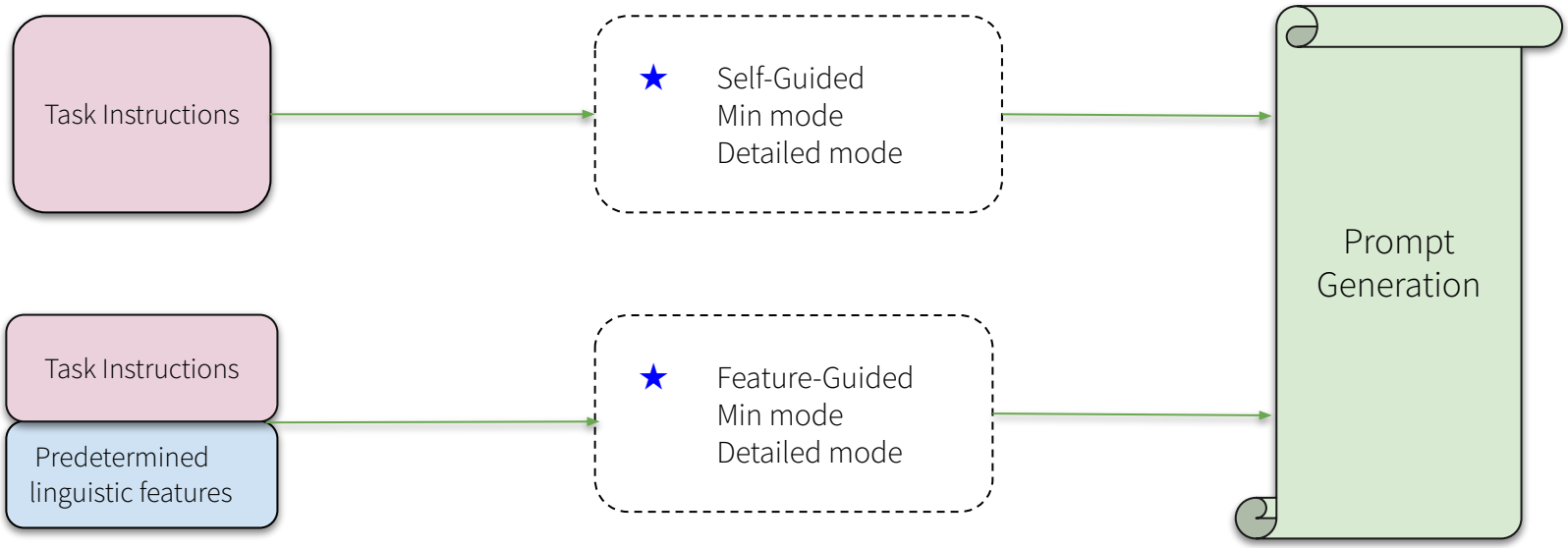
Findings of the 2019 Conference on Machine Translation (WMT19)
(Freitag et al., 2019 WMT)

Rewriting Process to Reduce Translationese

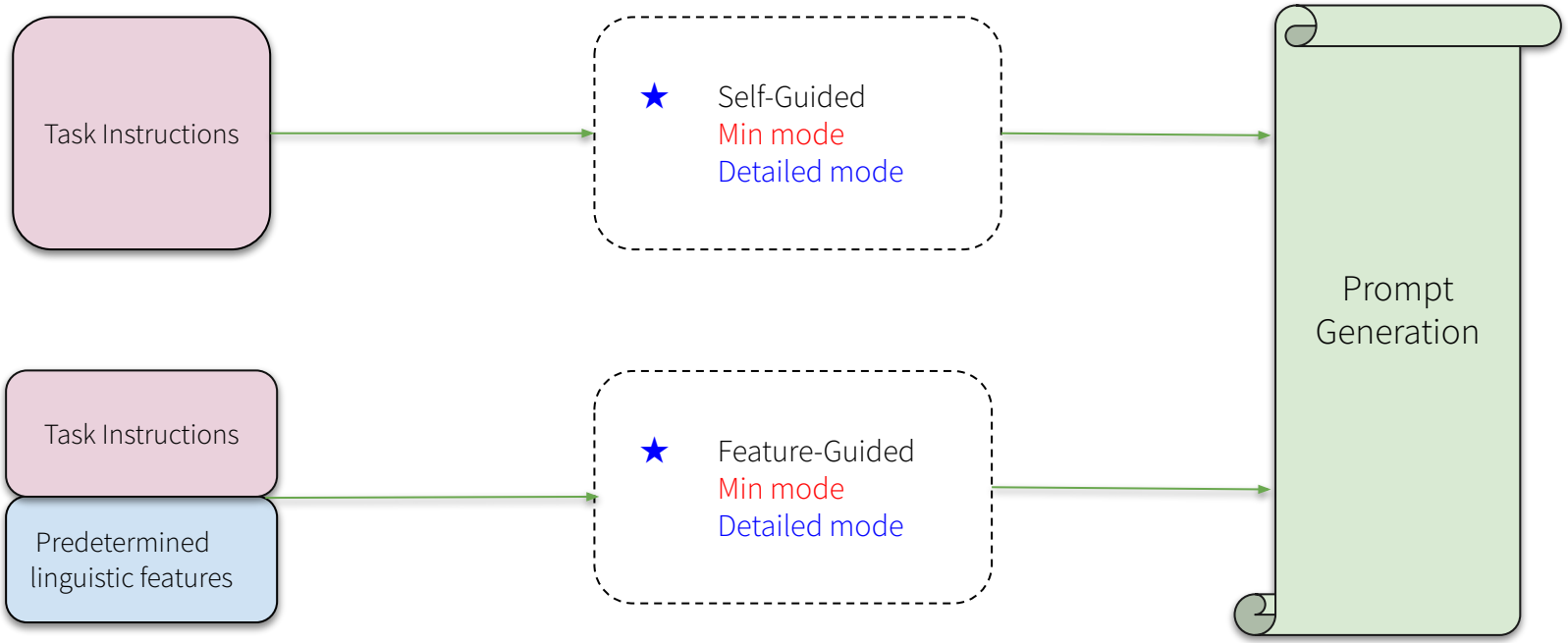
What details do prompts need to effectively reduce translationese in human-translated text?

Rewriting Process to Reduce Translationese

What details do prompts need to effectively reduce translationese in human-translated text?



Prompt Formulation



Prompt design: Self-Guided [Min]

Task presented in simple language; relies on the model's discretion without using technical terms related to translationese

Prompt = Source + Human Translation +Minimal Task Definition

Detailed Task Definition: *Given the original text in ${lang1}$, your task is to rewrite its human translation into ${lang2}$ in a more natural way if necessary.*
Revise the translation to sound natural if needed, otherwise return it unchanged.

Prompt design: Self-Guided [Detailed]

Provides a concise description of translationese, focusing on typical indicators identified in translation studies

Prompt = Source + Human Translation +Detailed Task Definition

Detailed Task Definition: *Given the original text in* ${lang1}$ *, task is to reduce translationese in human translation into* ${lang2}$ *by rewriting it in a more natural way where possible.* Translationese refers to any regular linguistic features in the translated texts that make them distinct from texts originally produced in the target language, outside the communicative situation of translation. … (+157 words)
Revise the translation to sound natural if needed, otherwise return it unchanged.

Handcrafted Translationese Indicators

- 58 Linguistic Features: morpho-syntactic features and text measures inspired by previous research on language-pair-specific translationese (Evert and Neumann, 2017; Kunilovskaya and Lapshinova-Koltunski, 2020), contrastive studies (König and Gast, 2007), and multilingual analysis (Hu and Kübler, 2021)

Evert, Stefan and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: a multivariate analysis for English and German. Empirical Translation Studies: New Methodological and Theoretical Traditions, 300:47– 80

Konig, Ekkehard and Volker Gast. 2007. Understanding English-German Contrasts. Erich Schmidt Verlag.

Kunilovskaya, Maria and Ekaterina Lapshinova- Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In the Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4102– 4112. The European Language Resources Association (ELRA).

Hu, Hai and Sandra Kübler. 2021. Investigating translated Chinese and its variants using machine learning. Natural Language Engineering, 27(3):339–372.

Handcrafted Translationese Indicators

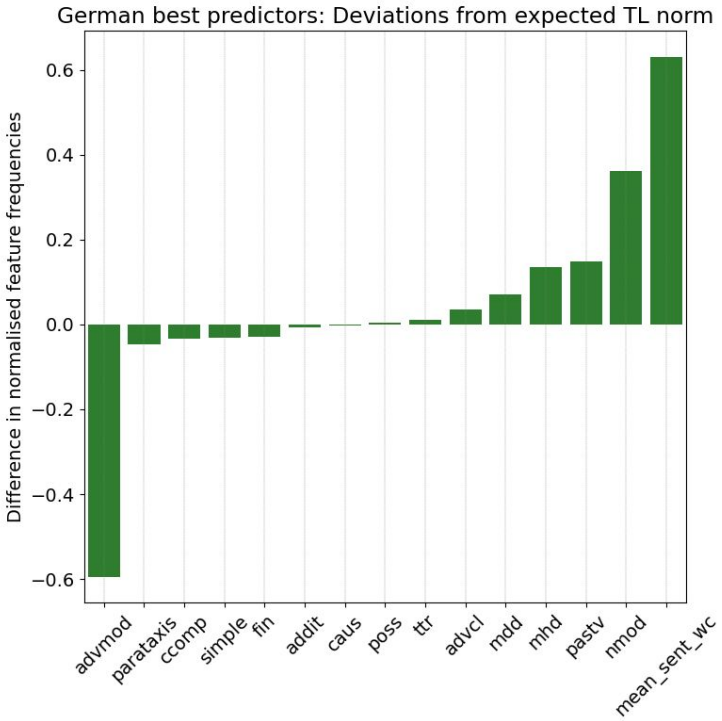
- 58 Linguistic Features: morpho-syntactic features and text measures

- Normalised frequencies of
  - selected word classes, esp. function words
  - syntactic dependencies
  - grammatical forms
  - Word order patterns

- Metrics
  - average word and sentence length
  - measures of lexical variety and density
  - mean hierarchical and mean dependency distances

Top Handcrafted Translationese Indicators

- ★ Instructions in the feature-guided minimal mode:
- ★ over-used features:
  - mean_sent_wc: 'Make sentences shorter.'
  - nmod: 'Avoid nominal modifiers of nouns such as prepositional and genitive case phrases.'
- ★ under-used features:
  - advmod: 'Use more adverbial modifiers.'
  - parataxis: 'Use explicit markers of discourse relations.'



direction of *deviations*

Prompt design: Feature-Guided [Min]

Tailored instructions for each text segment following the pre-compiled instructions

Prompt = Source + Human Translation + Task Definition + Min Instruction

Task definition: *Given the original text in ${lang1}$, your task is to rewrite its human translation into ${lang2}$ in a more natural way if necessary.*

Instructions: Revise this translation following the instructions:
Use pronouns instead of nouns as verbal arguments where possible.
Avoid constructions with indirect objects.

Prompt design: Feature-Guided [Detailed]

Detailed explanations of relevant linguistic concepts. Examples specific to the target language, when available.

Prompt = Source + Human Translation + Task Definition + Detailed Instruction

Task definition: *Given the original text in* ${lang1}$ *, your task is to reduce translationese in human translation into* ${lang2}$ *in a more natural way, less translated way.* Translationese refers to any properties of translations that make them statistically distinct from texts originally produced in the target language.

Instructions: Revise this translation following the instructions which reflect deviations of this segment from the expected target language norm: Use pronouns instead of nouns or proper names as verbal arguments where possible. Avoid constructions with indirect objects. An indirect object of a verb is any nominal phrase that is an obligatory argument of the verb but is not its subject or direct object. The prototypical example is the recipient (dem Kind) with verbs of exchange: Die Frau gibt dem Kind einen Apfel.

Experimental setup: Data

Europarl-UdS: transcripts of speeches & their written translation from EU Parliament website for EN<>DE

- ★ Entire Corpus: Balanced across translation directions (1500 random document pairs)
- ★ Contrastive Subset: ~ 200 documents in each TL that concentrated the translationese related phenomena

Entire corpus

		docs	segs	tokens
DE	original	1500	38,305	967,385
	translated	1500	36,078	924,919
EN	original	1500	36,078	927,045
	translated	1500	38,305	1,060,295

Classification with 58 feats.

		segs	tokens	seg_len ± std
DE	original	1908	59,942	31.4±17.6
	translated	1934	57,492	29.7±14.1
EN	original	1987	55,128	27.7±13.0
	translated	1919	65,065	33.9±19.6

Contrastive subset

Binary Classification Results

The main evaluation method for mitigating translationese is a segment-level translationese classification task, where lower scores compared to the baseline indicate better effectiveness.

		Baseline	Self-guided		Feature-guided	
			Min	Detail	Min	Detail
DE	15	81.06	-0.27	-1.01	**-2.39**	-2.21
	58	81.51	0.53	**-0.56**	0.06	-0.28
EN	15	75.60	-2.70	-4.10	-3.18	**-7.63**
	58	78.30	-1.40	-1.61	-1.61	**-4.07**

- The feature set effectively distinguishes human translations at the segment level

Binary Classification Results

Originals versus Translated → Yes or No?

(F1) lower is better

		Baseline	Self-guided		Feature-guided	
			Min	Detail	Min	Detail
DE	15	81.06	-0.27	-1.01	**-2.39**	-2.21
	58	81.51	0.53	**-0.56**	0.06	-0.28
EN	15	75.60	-2.70	-4.10	-3.18	**-7.63**
	58	78.30	-1.40	-1.61	-1.61	**-4.07**

- Both feature-guided min (15 features) and detail (58 features) are most effective

Binary Classification Results

Originals versus Translated → Yes or No?

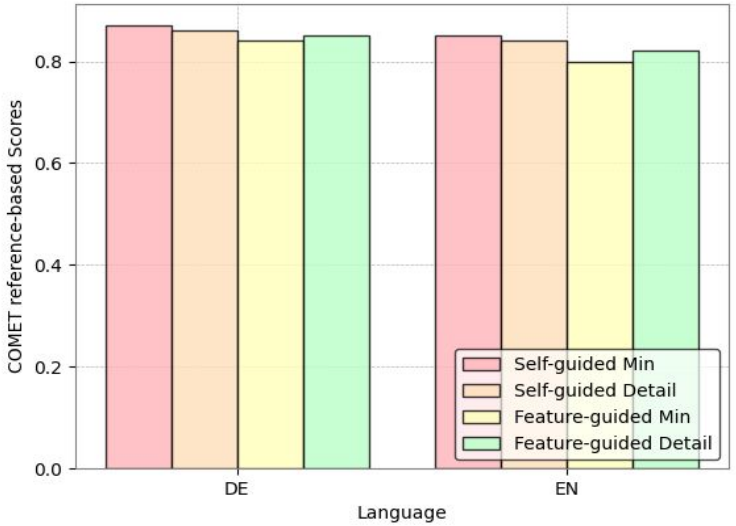
(F1) lower is better

		Baseline	Self-guided		Feature-guided	
			Min	Detail	Min	Detail
DE	15	81.06	-0.27	-1.01	**-2.39**	-2.21
	58	81.51	0.53	**-0.56**	0.06	-0.28
EN	15	75.60	-2.70	-4.10	-3.18	**-7.63**
	58	78.30	-1.40	-1.61	-1.61	**-4.07**

- feature-guided min (15 features) and self-guided detail (58 features) are most effective

Content Preservation

Evaluating the quality of GPT-4 outputs involves assessing how well they preserve the meaning of input translations using COMET (Rei et al., 2022)



High Level of

Content Preservation

Rei, Ricardo, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Manual Analysis

To rate translation accuracy and fluency on a 1- 6 scale for each output mode, with higher scores indicating better performance

		Rewriting Setups			
		Self-guided		Feature-guided	
		Min	Detail	Min	Detail
DE	A	5.8	5.8	5.1	5.4
	F	5.7	5.6	5.4	5.4
EN	A	5.7	5.9	5.2	5.4
	F	6	5.9	5.6	5.8

- Self-guided modes rated higher in accuracy and fluency than feature-guided modes for both English and German.

Summary

- We leverage the generative capabilities of GPT4 to reduce translationese-related differences between translated and non-translated texts
  - Our experimental approach involves two distinct prompting strategies: self-guided and feature-guided. These strategies vary in terms of the model's decision-making autonomy and the level of linguistic instruction provided
- Our findings demonstrate that GPT4 benefits from linguistic instructions and successfully mitigates translation artefacts in human translations, making them less distinguishable from non-translations
- Our best results for classification were with feature-guided instructions based on detailed linguistic descriptions
- More experiments are detailed in the paper

Summary (cont.)

Mitigating Translationese with GPT-4: Strategies and Performance

Maria Kunilovskaya[1], **Koel Dutta Chowdhury**[1], **Heike Przybyl** [1], **Cristina España-Bonet**[2], and **Josef van Genabith**[1,2]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]German Research Center for Artificial Intelligence (DFKI)
maria.kunilovskaya@uni-saarland.de

Abstract

Translations differ in systematic ways from texts originally authored in the same language. These differences, collectively known as translationese, can pose challenges in cross-lingual natural language processing: models trained or tested on translated input might struggle when presented with non-translated language. Translationese mitigation can alleviate this problem. This study investigates the generative capacities of GPT-4 to reduce translationese in human-translated texts. The task is framed as a rewriting process aimed at modified translations indistinguishable from the original text in the target language. Our focus is on prompt engineering that tests the utility of linguistic knowledge as part of the instruction for GPT-4. Through a series of prompt design experiments, we show that GPT4-generated revisions are more similar to originals in the target language when the prompts incorporate specific linguistic instructions instead of relying solely on the model's internal knowledge. Furthermore, we release the segment-aligned bidirectional German–English data built from the Europarl corpus that underpins this study.

1 Introduction

the outcomes of various cross-lingual tasks, potentially leading to biased results and decreased or artificially inflated performance, especially in evaluating machine translation (MT) models (Zhang and Toral, 2019; Graham et al., 2020), but also in the natural language inference tasks when using translated datasets and cross-lingual transfer scenarios (Artetxe et al., 2020). While translationese is viewed as an inalienable property of translated language, preferences may lean toward translation variants that are closer to target language patterns provided that the meaning and usefulness of the message in the source language (SL) are retained. The task of reducing translationese by making translations less deviant from the originally authored text in the target language (TL) is a newly recognised and relevant NLP problem. At the same time, only a few studies actively address it, including Dutta Chowdhury et al. (2022) who remove translation bias in latent representation space, as well as Jalota et al. (2023) and Wein and Schneider (2024), debiasing translations at the surface text level.

Our work is the first to explore the utility of linguistically informed prompts to harness the generative capabilities of large language models (LLMs) in the task of translationese mitigation. This approach is inspired by the successful application of LLMs to a range of text adaptation tasks including simplification (Feng et al., 2023), style transfer (Suzgun et al., 2022; Reif et al., 2022), and

Thank you All!

Summary (cont.)

Mitigating Translationese with GPT-4: Strategies and Performance

Maria Kunilovskaya[1], **Koel Dutta Chowdhury**[1], **Heike Przybyl** [1], **Cristina España-Bonet**[2], and **Josef van Genabith**[1,2]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]German Research Center for Artificial Intelligence (DFKI)
maria.kunilovskaya@uni-saarland.de

Abstract

Translations differ in systematic ways from texts originally authored in the same language. These differences, collectively known as translationese, can pose challenges in cross-lingual natural language processing: models trained or tested on translated input might struggle when presented with non-translated language. Translationese mitigation can alleviate this problem. This study investigates the generative capacities of GPT-4 to reduce translationese in human-translated texts. The task is framed as a rewriting process aimed at modified translations indistinguishable from the original text in the target language. Our focus is on prompt engineering that tests the utility of linguistic knowledge as part of the instruction for GPT-4. Through a series of prompt design experiments, we show that GPT4-generated revisions are more similar to originals in the target language when the prompts incorporate specific linguistic instructions instead of relying solely on the model's internal knowledge. Furthermore, we release the segment-aligned bidirectional German–English data built from the Europarl corpus that underpins this study.

1 Introduction

the outcomes of various cross-lingual tasks, potentially leading to biased results and decreased or artificially inflated performance, especially in evaluating machine translation (MT) models (Zhang and Toral, 2019; Graham et al., 2020), but also in the natural language inference tasks when using translated datasets and cross-lingual transfer scenarios (Artetxe et al., 2020). While translationese is viewed as an inalienable property of translated language, preferences may lean toward translation variants that are closer to target language patterns provided that the meaning and usefulness of the message in the source language (SL) are retained. The task of reducing translationese by making translations less deviant from the originally authored text in the target language (TL) is a newly recognised and relevant NLP problem. At the same time, only a few studies actively address it, including Dutta Chowdhury et al. (2022) who remove translation bias in latent representation space, as well as Jalota et al. (2023) and Wein and Schneider (2024), debiasing translations at the surface text level.

Our work is the first to explore the utility of linguistically informed prompts to harness the generative capabilities of large language models (LLMs) in the task of translationese mitigation. This approach is inspired by the successful application of LLMs to a range of text adaptation tasks including simplification (Feng et al., 2023), style transfer (Suzgun et al., 2022; Reif et al., 2022), and

Prompt design: Feature-Guided [Detailed]

Approach: Detailed explanations of relevant linguistic concepts. Examples specific to the target language, provided where possible.

Prompt template:

Task definition: *Your task is to reduce translationese in a human translation by re-writing it in a more natural, less translated way.* Translationese refers to any properties of translations that make them statistically distinct from texts originally produced in the target language.

Here is an original text in ${lang1}: ```${source}```

This is its human translation into ${lang2}: ```${human translation}```

Instructions: Revise this translation following the instructions which reflect deviations of this segment from the expected target language norm:

Use pronouns instead of nouns or proper names as verbal arguments where possible.
Avoid constructions with indirect objects. An indirect object of a verb is any nominal phrase that is an obligatory argument of the verb but is not its subject or direct object. The prototypical example is the recipient (dem Kind) with verbs of exchange: Die Frau gibt dem Kind einen Apfel.

Do not add any meta-phrases or quotation marks. Do not copy the original text.

Prompt design: Self-Guided [Min]

Approach: The task is presented in simple language, and relies on the model's discretion to address the issue without using technical terms related to translationese.

Prompt template:

Task definition: *Your task is to rewrite a human translation in a more natural way if necessary.*
Here is an original text in ${lang1}: ```${source}.```
This is its human translation into ${lang2}: ```${human translation}```

Instructions: If this translation can be revised to sound more like a text originally produced in the target language, return a revised version.
If this translation sounds natural enough, return the input translation.
Do not add any meta-phrases or quotation marks.
Do not copy the original text.

Prompt design: Self-Guided [Detailed]

Approach: This prompt provides a concise description of translationese, focusing on typical indicators identified in translation studies

Prompt template:

Task definition: *Your task is to reduce translationese in human translation by re-writing it in a more natural way where possible.* Translationese refers to any regular linguistic features in the translated texts that make them distinct from texts originally produced in the target language, outside the communicative situation of translation. These features are typically detected by statistical analysis and are explained by the specificity of the translation process. … (+140 words)

Here is an original text in ${lang1}: ```${source}```

This is its human translation into ${lang2}: ```${human translation} ```

Instructions: If you can detect any translationese deviations in this translation, revise this translation to make it sound less translated and return the revised version.

If no translationese is detected, return the input translation.

Do not add any meta-phrases or quotation marks. Do not copy the original text.

Prompt design: Feature-Guided [Min]

Approach: Tailored instructions for each text segment following the pre-compiled instructions

Prompt template:

Task definition: *Your task is to re-write a human translation in a more natural way.*
Here is an original text in ${lang1}: ```${source}```
This is its human translation into ${lang2}: ```${human translation}```

Instructions: Revise this translation following the instructions:
Use pronouns instead of nouns as verbal arguments where possible.
Avoid constructions with indirect objects.
Do not add any meta-phrases or quotation marks.
Do not copy the original text.

Qualitative Analysis

- Source: Does he regard the committee whose advice he rejects as incompetent, unable to judge what is good for animals or dangerous for human health?
- Human Translation: Hält er den Ausschuß, dessen Empfehlung er ablehnt, für nicht kompetent, für unfähig zu entscheiden, was für die Tiere gut oder für die menschliche Gesundheit gefährlich ist?
- Self-detailed: & Hält er den Ausschuss, dessen Ratschläge er ablehnt, für inkompetent und unfähig zu beurteilen, was gut für Tiere ist oder dem menschlichen Wohlbefinden schaden könnte?
- Feature-min: Betrachtet er den Ausschuss, dessen Rat er entschieden ablehnt, tatsächlich als inkompetent, als gänzlich unfähig zu beurteilen, was ausgerechnet für Tiere gut oder potenziell für die menschliche Gesundheit schädlich ist?
- Feature-detailed: Hält er tatsächlich den Ausschuß, dessen Empfehlung er konsequent ablehnt, für gänzlich inkompetent, für absolut unfähig, um zu beurteilen, was letztendlich für die Tiere gut oder womöglich für die menschliche Gesundheit gefährlich ist?

Mitigation of Translation Artefacts

- Translated texts form a “dialect” of the target language, reflecting the artifacts of the translation process, as well as traces of the source language. This “dialect” is known as *translationese*.
- Impacts Downstream Application
  - Confounds MT Evaluation
  - Stylistic deviations from originally-authored texts, when present in training data
  - Cross-Lingual Scenarios: Leads to artificially inflated performance on translated test sets

We propose a new technique: employing the generative capabilities of large language models (LLMs) to mitigate translationese

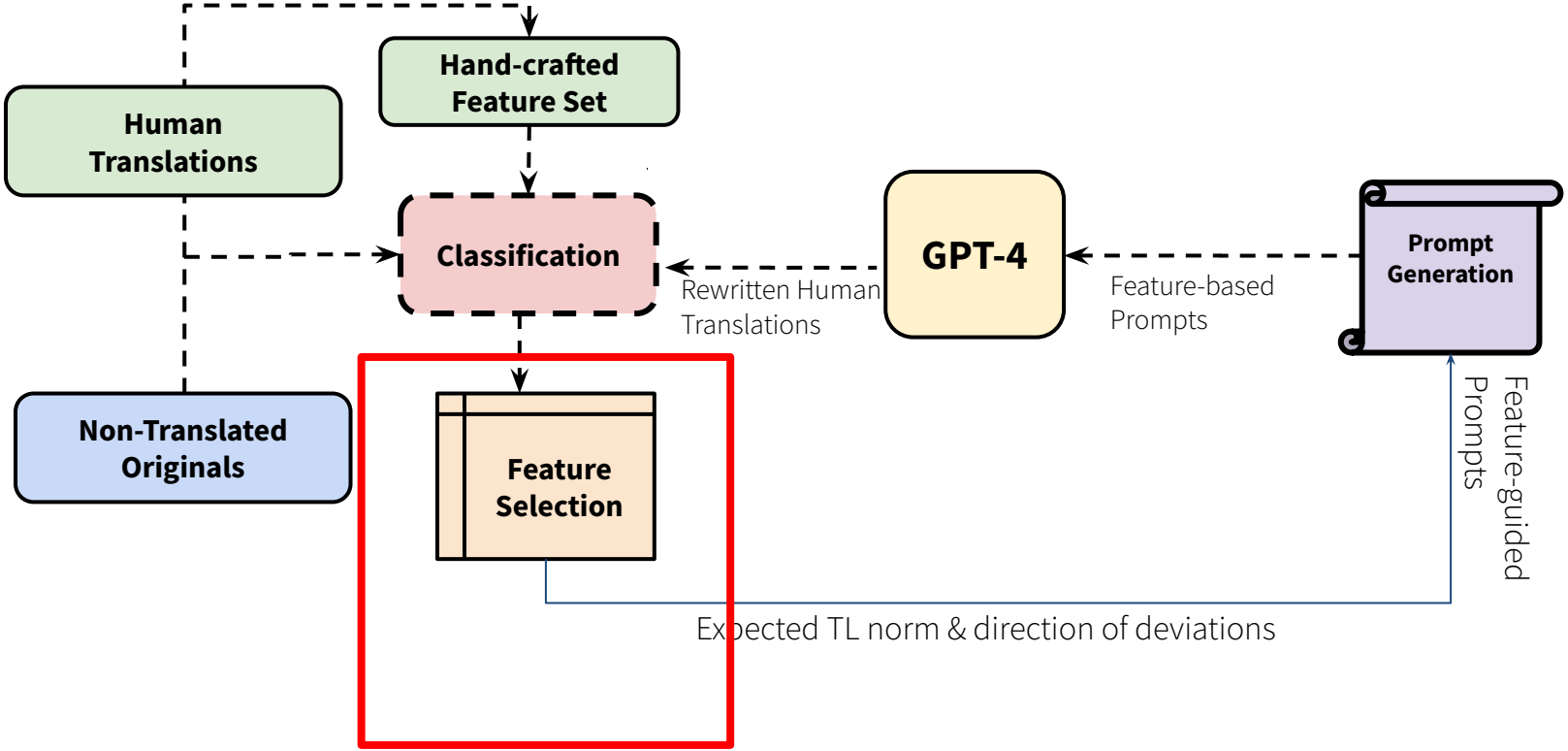
Existing Approaches to Translationese Mitigation

- Tagged-informed training (original and translationese tags) [1,2,3]
- Representational gap reduction in Latent Space:
  - Learning an original-to-translationese projection function [4]
  - Applying a debiasing approach to remove translationese from latent space [5]
- Intermediate Abstract Meaning Representation [6]
- *machine translation-based style transfer*

We propose a novel approach: leveraging the generative capabilities of large language models (LLMs) to mitigate translationese

1. Caswell, Isaac et al. “Tagged Back-Translation.” Conference on Machine Translation (2019).
2. Marie, Benjamin et al. “Tagged Back-translation Revisited: Why Does It Really Work?” Annual Meeting of the Association for Computational Linguistics (2020).
3. Riley, Parker, et al. "Translationese as a Language in “Multilingual” NMT." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.
4. Yu, S. et al. “Translate-Train Embracing Translationese Artifacts.” Annual Meeting of the Association for Computational Linguistics (2022).
5. Dutta Chowdhury, Koel et al. “Towards Debiasing Translation Artifacts.” North American Chapter of the Association for Computational Linguistics (2022).
6. Wein, Shira, and Nathan Schneider. "Translationese Reduction using Abstract Meaning Representation." *arXiv preprint arXiv:2304 .11501* (2023).

Overview of the Pipeline



Qualitative Analysis

- Source: Der Ruhm Portugals ist gemehrt worden durch diese Ratspräsidentschaft, vor allem durch drei Dinge, die Sie selbst zitiert haben.
- Human Translation: Portugal’s reputation has grown with this Presidency, specifically thanks to three things, which you yourself have mentioned.
- Self-detailed: Portugal's reputation has been enhanced by this Presidency, particularly due to three things you've highlighted.
- Feature-detailed: The prestige of Portugal has expanded due to this Presidency, attributed to three aspects you've referenced.

Experimental setup: Data

Europarl-UdS: transcripts of speeches (adapted for reading) and their written translation from European Parliament website (up to 2018 corpus extracted using Martinez's pipeline) for EN<>DE

+ Automatically aligned at sentence level using domain-specific dictionaries (+assessment of alignment quality)
+ Balanced across translation directions (1500 random document pairs)
+ Filtered by sentence length (450 or more tokens in the source)

Entire corpus

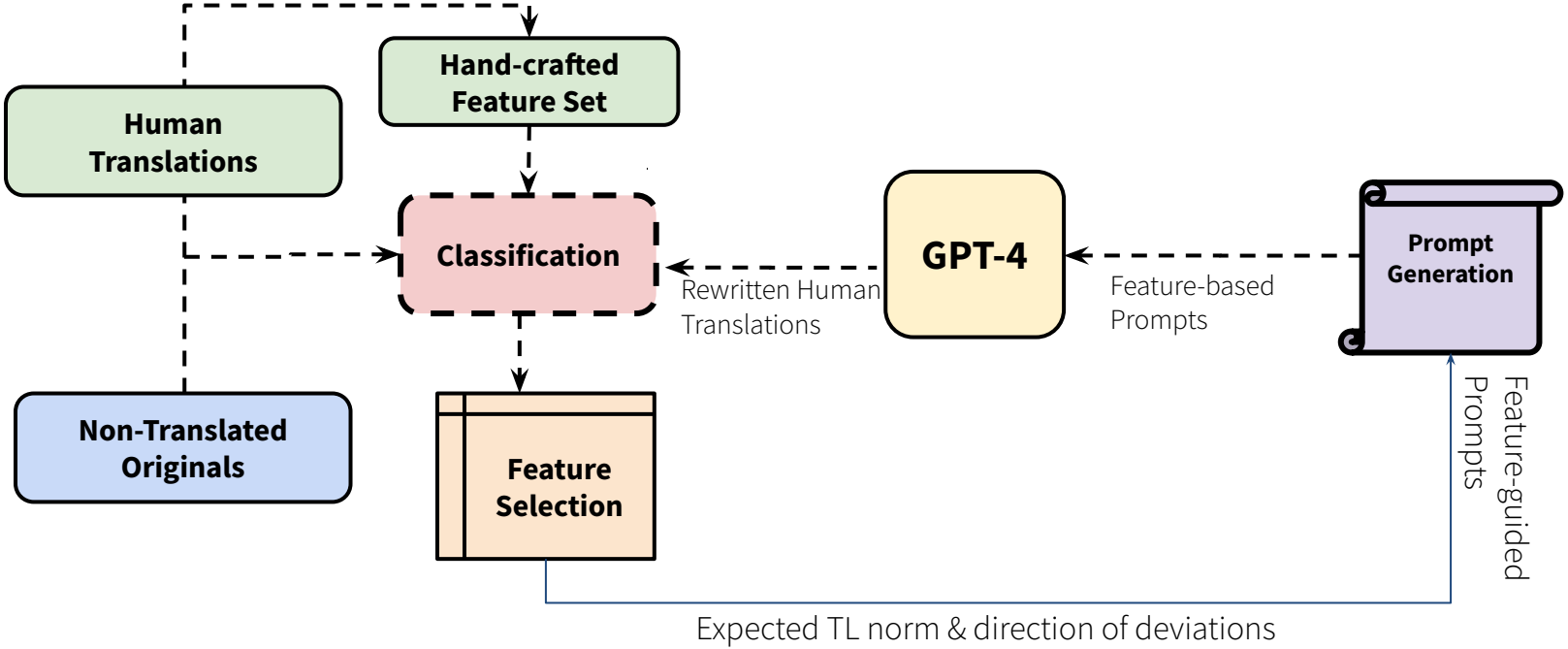
		docs	segs	tokens
DE	original	1500	38,305	967,385
	translated	1500	36,078	924,919
EN	original	1500	36,078	927,045
	translated	1500	38,305	1,060,295

Download: https://zenodo.org/records/11127626

Contrastive subset

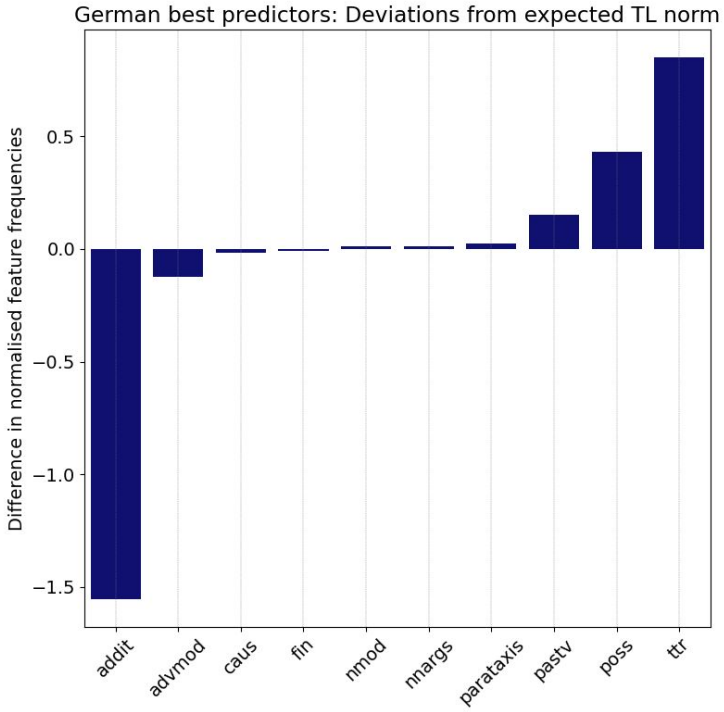
		segs	tokens	seg_len ± std
DE	original	1908	59,942	31.4±17.6
	translated	1934	57,492	29.7±14.1
EN	original	1987	55,128	27.7±13.0
	translated	1919	65,065	33.9±19.6

Overview of the Pipeline

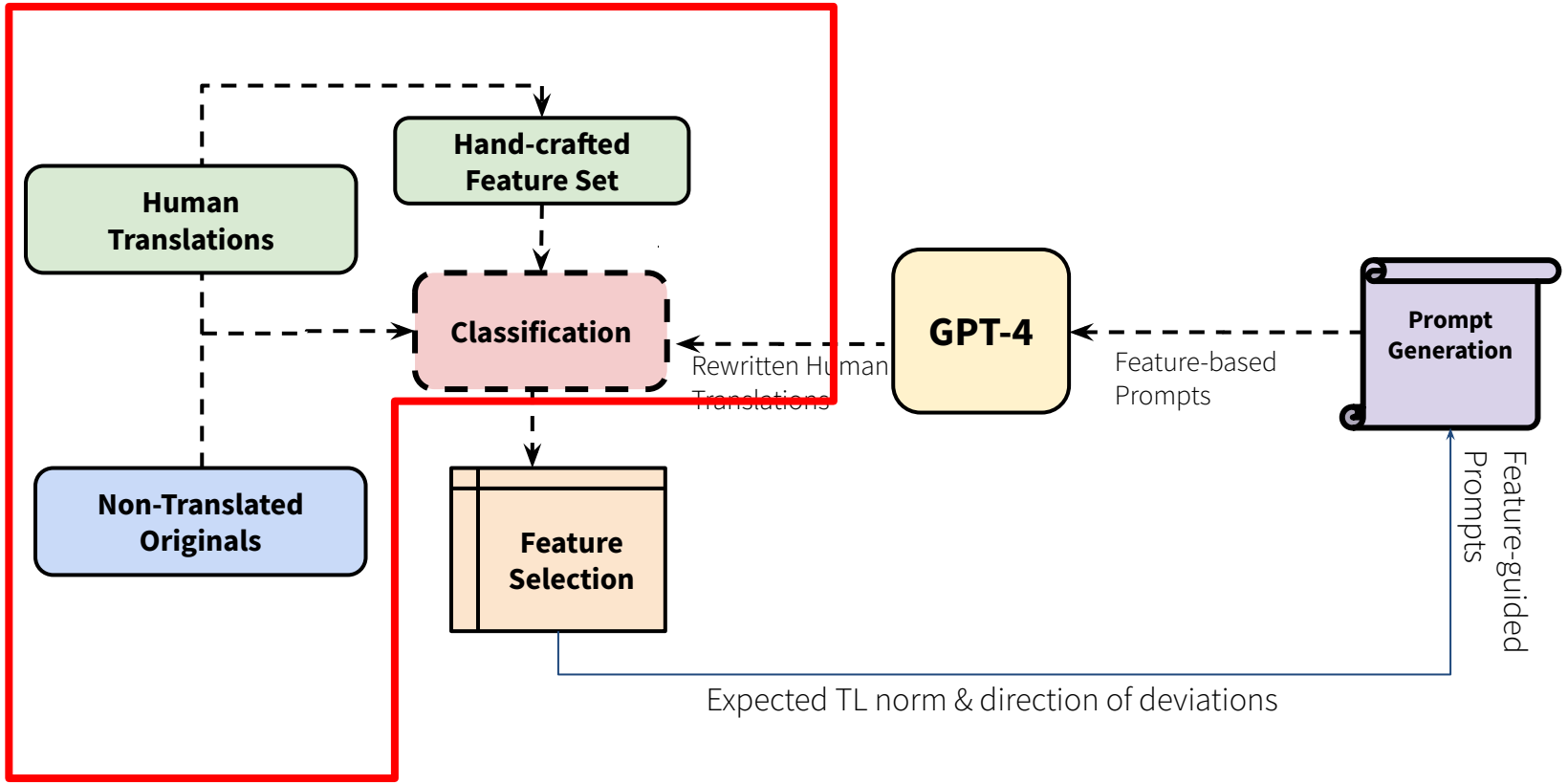


Top Handcrafted Translationese Indicators

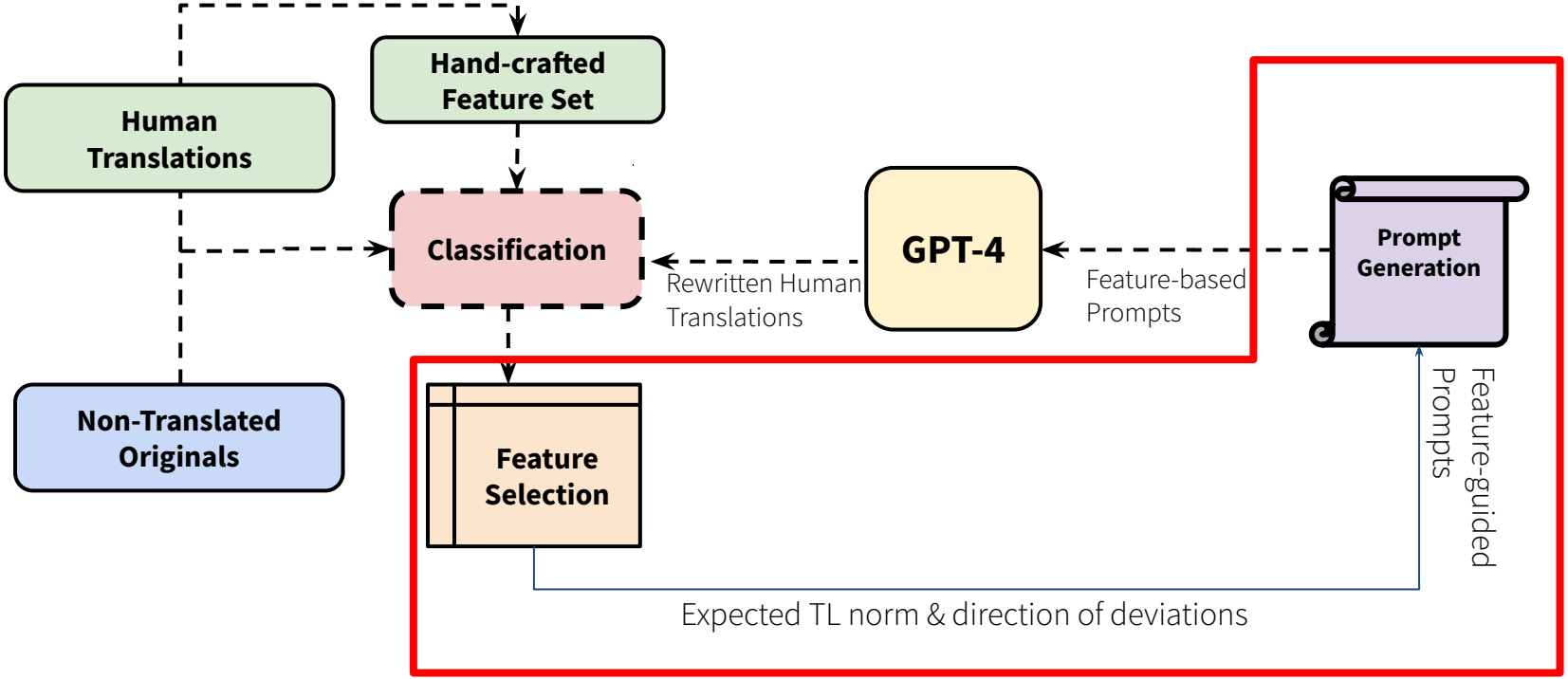
- Feature specified instructions
  - acl: ‘allow more attributive clauses’ or ‘avoid attributive clauses’,
  - Parataxis: ‘avoid explicit connectives’ or ‘use explicit markers of discourse relations’,
  - ttr:‘avoid lexical repetitions’ or ‘rely on more frequent words’



Overview of the Pipeline

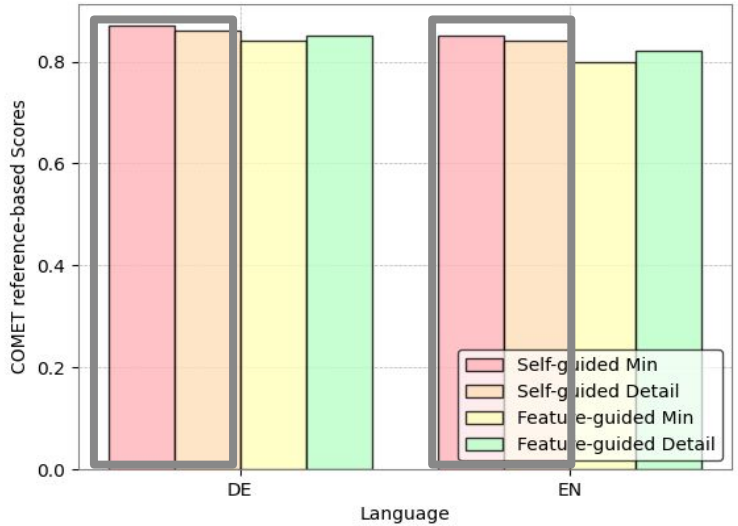


Overview of the Pipeline



Content Preservation

Task: Evaluating the quality of GPT-4 outputs involves assessing how well they preserve the meaning of input translations using COMET (Rei et al., 2022)



High Level of

Content Preservation

Rei, Ricardo, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.