C A N C E R: Corpus for Accurate Non-English Cancer-related Educational Resources

Anika Harju[1] **Asma Shakeel**[2] **Tiantian He**[3] **Tianqi Xu**[3] **Aaro Harju**[4]
[1] University of Technology Sydney, Australia
[2] School of Electrical Engineering and Computer Science, NUST, Pakistan
[3] IT University of Copenhagen, Denmark
[4] Independent Researcher
Anika.Harju@.uts.edu.au, ashakeel.msee18seecs@seecs.edu.pk,
{tihe,tixu}@itu.dk, aaro.m.harju@outlook.com

Abstract

Improving the quality of cancer terminology through Machine Translation (MT) in non-English languages remains an under-researched area despite its critical role in supporting self-management and advancing multilingual patient education. Existing computational tools encounter significant limitations in accurately translating cancer terminologies, particularly for low-resource languages, primarily due to data scarcity and morphological complexity. To address the gap, we introduce a dedicated terminology resource —Corpus for Accurate Non-English Cancer-related Educational Resources (C A N C E R), a manually annotated dataset in Finnish (FI), Chinese (ZH), and Urdu (UR), curated from publicly available existing English (EN) data. We also examine the impact of data quality versus quantity and compare the performance of the Opus-mt-en-fi, Opus-mt-en-zh, and Opus-mt-en-ur models with the SMaLL-100 multilingual MT model. We assess translation quality using automatic and human evaluation. Results demonstrated that high-quality parallel data, though sparse, combined with fine-tuning, substantially improved the translation of cancer terminology across both high and low-resource language pairs, positioning the C A N C E R corpus as a foundational resource for improving multilingual patient education.[1]

1 Introduction

Cancer remains a major global health challenge, representing one of the leading causes of death worldwide (Bray et al., 2021). Patient education is critical for understanding the cancer diagnosis and undergoing the intensive treatment (Cai et al., 2023). There is a significant demand for simplifying complex cancer terminology through Machine Translation (MT) in patient education materials to improve health literacy (Oniani et al., 2023). The persistent research gap impedes effective cancer patient education and increases the risk of misdiagnosis and adverse outcomes (Kasperė et al., 2023). Moreover, the World Health Organization International Classification of Diseases recommends the translation of medical terminology into other languages to enhance accessibility, as codes and classifications containing the ontologies are primarily in English (EN) (Harrison et al., 2021). Consequently, the accurate translation of medical terminology, particularly for diseases such as cancer, is critical for advancing cancer patient education and self-management (McCorkle et al., 2011) in support of patients with limited proficiency in the native language where they reside (Castilla et al., 2005; Lovis et al., 1998).

Despite the high proficiency of state-of-the-art (SOTA) Neural Machine Translation (NMT) models (Dabre et al., 2020; Wang et al., 2023), MT of medical terminology has fallen short (Nayak et al., 2023). Even with various fine-tuning approaches, NMT models still struggle to translate medical terminology accurately (Nayak et al., 2020). One approach to mitigate the issue is to utilize a high-quality parallel dataset for MT training (de Gibert Bonet et al., 2022). However, annotated parallel medical data remain scarce — particularly in the cancer domain (Ma et al., 2020). Furthermore, the computational demands associated with implementing MT on SOTA models are costly (Nayak et al., 2023; Park et al., 2021; Zhang et al., 2023).

In this paper, we focus on fine-tuning three NMT models (Opus-mt-en-fi, Opus-mt-en-zh, and Opus-mt-en-ur) (Tiedemann and de Gibert, 2023) and a multilingual MT model (SMaLL-100) (Mohammadshahi et al., 2022) using manually annotated training data derived from EN segments of the public English-Chinese Cancer Parallel Corpus (ECCParaCorp) (Ma et al., 2020) to construct new EN-to-FI and EN-to-UR parallel corpora and extend language coverage of the existing EN-to-

[1] Available at: C A N C E R Corpus

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 197–209, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-022

Annotated data	# Pairs
In-domain	
EN-to-FI	1,494
EN-to-ZH	1,494
EN-to-UR	1,494
Out-of-domain	
EN-to-FI	291
EN-to-ZH	291
EN-to-UR	291
Total	5,355

Table 1: Annotated cancer terminology parallel data

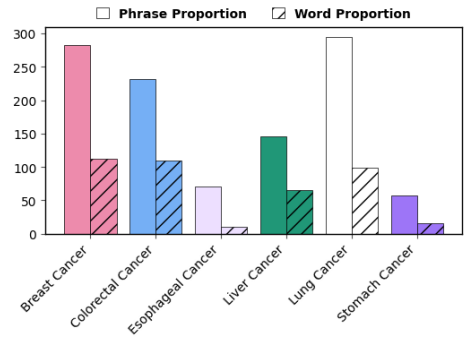

Figure 1: Categories of terminology data in cancer advocacy colors

ZH language pair. We assess translation quality using automatic evaluation metrics (Papineni et al., 2002; Popović, 2015; Rei et al., 2020) and human evaluation (Escribe, 2019). We also evaluate generalization using human evaluation on three manually annotated parallel datasets (EN–FI, EN–ZH, and EN–UR) curated from the public glossary on the Peter MacCallum Cancer Centre website (MacCallum, 2024).

Our paper focuses on improving the translation quality of cancer terminologies in two high-resource languages (FI and ZH) and a low-resource language (UR) to advance cancer patient education and bridge language challenges to support improved self-management (Lovis et al., 1998).

Our contributions can be summarized as follows:

- Creation of C A N C E R, a manually annotated corpus, to advance cancer patient education and self-management in EN-to-FI, EN-to-ZH, and EN-to-UR language pairs.
- Adaptation of the Opus-mt-en-fi, Opus-mt-en-zh, Opus-mt-en-ur, and the SMaLL-100 multilingual MT model, through fine-tuning to improve the translation quality of cancer terminologies.
- In-depth analysis of automatic performance metrics, including human evaluation by medical practitioners and native FI, ZH, and UR speakers provided insights into the translation quality of the cancer terminologies.

2 Data

In the first data acquisition step, we collected EN data from Ma et al. (2020) cancer corpus, which includes cancer terminologies (411 words and 1,083 phrases) (Table 1) related to cancer prevention, screening, diagnosis, and treatment. Using the existing EN source data, we manually annotated the FI and UR references to create two parallel datasets (EN-to-FI and EN-to-UR) while extending language coverage with the EN-to-ZH pair for the training, development, and test splits (Ma et al., 2020). We excluded sentence-level data to focus exclusively on terminology-level translation. The C A N C E R corpus includes data in six categories: Breast Cancer, Colorectal Cancer, Esophageal Cancer, Liver Cancer, Lung Cancer, and Stomach Cancer (Figure 1). In the second step, we compiled EN data from MacCallum (2024) online glossary that covers commonly used cancer-related terminologies (182 words and 109 phrases) (Table 1) from A-to-Z during diagnosis and treatment, and manually annotated FI, ZH, and UR references to the source data to create three out-of-domain datasets (EN-to-FI, EN-to-ZH, and EN-to-UR) to assess generalization using human evaluation (Escribe, 2019). We annotated our five parallel cancer terminology datasets, leveraging the expertise of medical practitioners and native FI, ZH, and UR speakers, effectively addressing the data scarcity gap (Lovis et al., 1998).

3 Related Work

Translating medical terminologies is a challenging task. The unique features of different languages, combined with the complexity of medical jargon and data scarcity, have further hindered efforts. (Ao and Acharya, 2021) Moreover, medical institutions have limited specialized health educators to support self-healthcare in chronic diseases such as cancer, particularly for non-native English speakers (Ugas et al., 2024). Existing studies have sought to improve translation quality across the medical domain, including work to enhance med-

ical terminology, without specific emphasis on the cancer field (Alam et al., 2021). Prior research included exploring the time frame required to translate newly introduced or revised medical terminology for evaluation by healthcare experts (Skianis et al., 2020). The Castilla et al. (2005) study also investigated automated evaluation of medical terminologies using the Unified Medical Language System to assess cross-lingual information clinical data extracted from Portuguese-language thoracic radiology reports. During human evaluation, the Kasperė et al. (2024) study, however, found that the translation of medical terminology from English to Lithuanian was of poor quality, concluding that MT should serve as a supplementary approach only. In contrast, the Herrera-Espejel and Rach (2023) study highlighted MT as a potential solution to bridge language barriers in public health communication that restrict access to essential information for culturally and linguistically diverse groups.

In our experiment, we rely on manually annotated domain-specific data (Table 1) and fine-tuning techniques to adapt the Opus-MT models and the SMaLL-100 multilingual MT model to the unique morphological characteristics of FI, ZH, and UR cancer terminologies to advance health education (Oniani et al., 2023), and support patients to overcome language barriers, particularly when taking prescribed medication and navigating digital platforms (Lorig and Holman, 2003; McCorkle et al., 2011).

4 Method

We denote, let $X = \{x_1, x_2, \ldots, x_N\}$ as the source language (EN) consisting of N medical terminologies. $Y = \{y_1, y_2, \ldots, y_N\}$ as its corresponding target-language (FI, ZH and UR). Each pair (x_i, y_i) constitutes a parallel medical terminology. The probability of translating the entire target sequence Y given the source sequence X can be approximated as:

$$P(Y \mid X; \theta) \approx \prod_{i=1}^{N} P(y_i \mid x_i; \theta)$$

OPUS-MT In the first stage of the experiment, we fine-tuned the Opus-MT models on annotated parallel training data (EN-FI, EN-ZH, and EN-UR). We utilized dynamic batching with the Hugging Face DataCollatorForSeq2Seq (Solanki and Khublani, 2024) and systematically optimized hyperparameters by experimenting with batch sizes 8, 16, and 32 (achieving the best performance with a batch size of 8) and a learning rate grid search (optimal rate: 6e-04) over three epochs (Appendix A). Label smoothing (probability = 0.1) was applied to enhance precision. We evaluated model performance using the bilingual evaluation understudy (BLEU) (Vaswani, 2017), CHaRacter-level F-score (CHRF) (Popović, 2015), and Cross-lingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020) metrics .

SMaLL-100 The second stage of the experiment involved prepending the EN language token (_en_) to the encoder input in the SMaLL-100 model to specify the source language explicitly. To prompt the decoder to generate translations in the correct target language, we added a beginning-of-sequence (BOS) token via the forced_bos_token_id parameter. We applied similar hyperparameter settings (Appendix A) as in the first experiment to ensure consistency across model comparisons, using an optimal learning rate of 7e-05 (achieving the best performance with a batch size of 8) (Fuady et al., 2024). Native speakers assessed the generated translations on the in-domain test data from both experiments. To evaluate generalization, we selected the models with the lowest validation loss and assessed translation quality on out-of-domain datasets using human evaluation (Escribe, 2019).

5 Results

OPUS-MT The models demonstrated varying levels of translation effectiveness across the EN-FI, EN-ZH, and EN-UR language pairs. The Opus-mt-en-fi model achieved the highest BLEU score (Table 2), suggesting robust translation quality. CHRF and COMET scores (Appendices B & C) were also consistently high, indicating strong alignment with reference translations at the character and semantic level. The stability highlighted the capacity of the Opus-mt-en-fi model to adapt to the intricate morphological structure of the FI language, reinforcing its suitability for the MT task. Similarly, the Opus-mt-en-zh model exhibited satisfactory performance across various configurations (Appendices B & C), highlighting the ability to understand the language patterns. However, performance dipped with the Opus-mt-en-ur model, as challenges persist in generalizing across

Batch Size	Opus-mt-en-fi			Opus-mt-en-zh			Opus-mt-en-ur			EN-FI			EN-ZH			EN-UR		
										SMaLL-100 Multilingual MT Model								
	BLEU	CHRF	COMET	BLEU	CHRF	COMET	BLEU	CHRF	COMET	BLEU	CHRF	COMET	BLEU	CHRF	COMET	BLEU	CHRF	COMET
Baseline																		
8	12.95	51.12	82.06	7.61	24.14	75.77	2.38	16.51	51.26	3.43	13.82	65.30	2.78	6.38	65.62	2.60	1.68	54.87
16	12.73	50.58	81.75	3.67	21.28	75.02	2.17	16.46	51.16	2.40	13.39	64.48	1.43	5.41	64.50	1.34	1.53	53.83
32	11.52	49.81	81.07	2.38	19.05	73.89	2.08	16.44	51.12	1.92	12.81	63.51	1.30	5.24	64.11	1.22	1.51	53.45
Fine-tuned																		
8	**58.25**	**75.22**	**92.24**	**41.28**	**48.46**	**86.96**	**28.60**	**47.20**	**68.30**	**54.40**	**73.13**	**88.04**	**40.92**	**48.03**	**85.06**	**44.93**	**66.53**	**80.03**
16	**57.37**	**74.24**	**91.62**	**44.48**	**53.15**	**86.28**	**27.46**	**47.98**	**68.48**	**54.35**	**73.18**	**88.00**	**41.12**	**48.81**	**84.82**	**45.43**	**65.30**	**79.94**
32	**57.96**	**75.93**	**92.01**	**43.12**	**53.57**	**86.80**	**30.25**	**49.62**	**70.68**	**53.62**	**72.41**	**87.63**	**07.18**	**26.28**	**75.18**	**06.12**	**29.91**	**63.56**

Table 2: Automatic evaluation metrics for the OPUS-MT models and the SMaLL-100 MT model

the unique linguistic structures of the UR language. The reduced scores (Appendices B & C) indicated the Opus-mt-en-ur model experienced difficulties in capturing the complexity of the UR language, likely due to distinct syntactic characteristics.

SMaLL-100 In contrast, the SMaLL-100 model demonstrated improved performance on the EN-UR language pair, surpassing the Opus-mt-en-ur at smaller batch sizes, suggesting better adaptability to the unique linguistic structures of UR. However, performance declined significantly at a batch size of 32, resulting in low scores. (Appendices B & C) The model exhibited performance trends similar to the OPUS-MT models across the EN-FI and EN-ZH language pairs. On the EN-FI pair, the model achieved competitive BLEU and CHRF scores, though slightly lower than the Opus-mt-en-fi model (Table 2). The SMaLL-100 model demonstrated comparable performance to Opus-mt-en-zh on smaller batch sizes, with only a slight decline in BLEU and COMET scores. Translation quality declined, however, on the EN-ZH pair at a batch size of 32. (Appendices B & C)

Based on the results (Table 2, Appendices B & C), we hypothesize that the Opus-MT models outperformed the SMaLL-100 model due to language-specific training, which enabled optimization and improved translation quality.

6 Analysis

Automatic Evaluation Overall, the Opus-mt-en-fi model demonstrated robust performance on the EN-FI language pair (Appendices B & D). The Opus-mt-en-fi achieved the highest BLEU scores (58.25, 57.37, and 57.97) on the MT task, closely followed by the SMaLL-100 model. Both models maintained strong consistency on the EN-FI language pair (Table 2). Similarly, the Opus-mt-en-zh model demonstrated satisfactory translation quality across all batch sizes. The SMaLL-100 model matched the performance stability at smaller batch sizes (8 and 16). However, performance declined at batch size 32 on the EN-ZH corpus, which showed a reduction in effectiveness and translation quality (Appendix B). Notably, the SMaLL-100 model demonstrated stronger performance than the Opus-mt-en-ur model at smaller batch sizes (8 and 16), which suggested the multilingual model was more effective in capturing the unique language patterns of the UR language. Performance declined significantly at batch size 32, mirroring patterns observed in the EN-ZH language pair. (Appendix B)

Language Pair	Correct (%)	Partially Correct (%)	Incorrect (%)
In-domain			
EN - FI	67.34	25.17	07.50
EN - ZH	45.85	06.83	47.32
EN - UR	26.57	60.78	12.65
Out-of-domain			
EN - FI	54.98	09.62	35.40
EN - ZH	20.27	06.19	73.54
EN - UR	06.19	21.99	71.82

Table 3: Percentage-based human evaluation across language pairs

Human Evaluation A qualitative analysis guided the human evaluation to determine whether the translations were correct, partially correct, or incorrect (Table 3). The evaluators observed multiple gold-standard translations (Appendices D, E & F) and some discrepancies (Appendix G) across the EN-FI, EN-ZH, and EN-UR language pairs, highlighting differences in generalization among the models. In a few cases, the human evaluators noticed that the models generated synonyms for some cancer terminologies, skipped translations, and produced grammatical errors (Appendix G).

Skipped Translations In some instances, no translation occurred across the language pairs, indicating limitations in the capacity of the Opus-MT and SMaLL-100 models to convert source references into the target language due to the unique morphological structure of each language.

For instance, the term *Topotecan* remained in its EN form, not matching the ZH reference. (Appendix G)

Grammatical Errors Punctuation and spacing errors occurred during the translation of some terminologies. While the models translated the cancer terminologies accurately, the generated output did not include the unique grammatical rule of the specific target language. (Appendix G)

Ambiguous Terms Some translations featured incorrect word order or introduced extraneous tokens, which distorted the intended meaning of the target reference. Additionally, an extra token generated during translation distorted the ZH reference for the term *Vancomycin-resistant Enterococcus*. Similarly, in the UR language pair, the term *advanced age* did not align with the target reference, reflecting a syntactic and semantic mismatch of the target language. (Appendix G)

7 Limitations

A significant limitation of the task was the size of the annotated corpus. The C A N C E R corpus included limited data in only three languages out of more than 7,000 spoken worldwide, which restricted the scope of the findings and applicability to broader multilingual contexts. While model performance was satisfactory overall, the data constraint likely contributed to the instability observed in the multilingual SMaLL-100 model at higher batch sizes, where translation quality degraded. Additionally, the UR language presented unique challenges due to its right-to-left script, which may have complicated the tokenization process. The limitations necessitate the need to expand the corpus and further experiment with optimizing techniques and models to improve translation quality across languages.

8 Conclusion and Future Work

In this paper, we took the first step towards advancing multilingual cancer patient education. The C A N C E R corpus serves as a benchmark resource for evaluating the translation of cancer terminology across languages. The findings inform efforts to improve multilingual cancer patient education, supporting non-native English speakers in understanding critical health information. We demonstrated that retraining on limited high-quality parallel data (Shin et al., 2020) can improve translation quality (Table 2). In future work, we aim to expand the C A N C E R corpus by incorporating a broader spectrum of low and high-resource languages and exploring varying techniques and NMT models to optimize performance, mainly in underrepresented languages.

Acknowledgements

We want to express our gratitude to Deputy Head of School (Research) Dr. Camille Dickson-Deane and Senior Lecturer Dr. Amara Atif of the University of Technology Sydney for the final review of the paper and Associate Professors Rob van der Goot and Christian Hardmeier of the IT University of Copenhagen for feedback on the initial version. We are also grateful to the medical practitioners who reviewed segments of the annotated data to verify the target references of some terminologies.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the wmt shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663.

Shuang Ao and Xeno Acharya. 2021. Learning ulmfit and self-distillation with calibration for medical dialogue system. *arXiv preprint arXiv:2107.09625*.

Freddie Bray, Mathieu Laversanne, Elisabete Weiderpass, and Isabelle Soerjomataram. 2021. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127(16):3029–3030.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz, and Hong Yu. 2023. PaniniQA: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

Andre Castilla, Alice Bacic, and Sergio Furuie. 2005. Machine translation on the medical domain: the role of bleu/nist and meteor in a controlled vocabulary setting. In *Proceedings of Machine Translation Summit X: Papers*, pages 47–54.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Ona de Gibert Bonet, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé, and Maite Melero. 2022. Quality versus quantity: Building catalan-english mt resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69.

Marie Escribe. 2019. Human evaluation of neural machine translation: The case of deep learning. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 36–46.

Muhammad Jauharul Fuady, Alvalen Shafelbilyunazra, Agusta Rakhmat Taufani, Yogi Dwi Mahandi, and Shofiyah Al Idrus. 2024. Hyperparameter optimization for transformer-based translation models on low-power devices. In *2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C)*, pages 631–636. IEEE.

James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. Icd-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21:1–10.

Paula Sofia Herrera-Espejel and Stefan Rach. 2023. The use of machine translation for outreach and health communication in epidemiology and public health: scoping review. *JMIR Public Health and Surveillance*, 9(1):e50814.

Ramunė Kasperė, Jurgita Mikelionienė, and Dalia Venckienė. 2023. Medical terminology issues: a feasibility study of machine translation in a low-resource language. *SKASE Journal of Translation and Interpretation*, 16(2):5–22.

Ramunė Kasperė, Jurgita Mikelionienė, and Dalia Venckienė. 2024. Medical terminology issues: a feasibility study of machine translation in a low-resource language. *SKASE Journal of Translation and Interpretation*, 1:5–22.

Kate R Lorig and Halsted R Holman. 2003. Self-management education: history, definition, outcomes, and mechanisms. *Annals of behavioral medicine*, 26(1):1–7.

Christian Lovis, Robert Baud, Anne-Marie Rassinoux, Pierre-André Michel, and Jean-Raoul Scherrer. 1998. Medical dictionaries for patient encoding systems: a methodology. *Artificial intelligence in medicine*, 14(1-2):201–214.

Hetong Ma, Feihong Yang, Jiansong Ren, Ni Li, Min Dai, Xuwen Wang, An Fang, Jiao Li, Qing Qian, and Jie He. 2020. Eccparacorp: a cross-lingual parallel corpus towards cancer education, dissemination and application. *BMC Medical Informatics and Decision Making*, 20:1–12.

Peter MacCallum. 2024. Everyday cancer words and terms: A to z. Victoria, Australia. Available at https://www.petermac.org/patients-and-carers/information-and-resources/a-z-of-everyday-cancer-words-and-terms.

Ruth McCorkle, Elizabeth Ercolano, Mark Lazenby, Dena Schulman-Green, Lynne S Schilling, Kate Lorig, and Edward H Wagner. 2011. Self-management: Enabling and empowering patients living with cancer as a chronic illness. *CA: a cancer journal for clinicians*, 61(1):50–62.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prashanth Nayak, Rejwanul Haque, and Andy Way. 2020. The adapt' s submissions to the wmt20 biomedical translation task.

Prashanth Nayak, John Kelleher, Rejwanul Haque, and Andy Way. 2023. Instance-based domain adaptation for improving terminology translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 222–234.

David Oniani, Sreekanth Sreekumar, Renuk DeAlmeida, Dinuk DeAlmeida, Vivian Hui, Young Ji Lee, Yiye Zhang, Leming Zhou, and Yanshan Wang. 2023. Toward improving health literacy in patient education materials with neural machine translation models. *AMIA Summits on Translational Science Proceedings*, 2023:418.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chanjun Park, Sugyeong Eo, Hyeonseok Moon, and Heui-Seok Lim. 2021. Should we find another model?: Improving neural machine translation performance with one-piece tokenization method without model modification. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies: Industry papers*, pages 97–104.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706.

Konstantinos Skianis, Yann Briand, and Florent Desgrippes. 2020. Evaluation of machine translation methods applied to medical terminologies. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 59–69, Online. Association for Computational Linguistics.

Shivam R Solanki and Drupad K Khublani. 2024. Generative large language models. In *Generative Artificial Intelligence: Exploring the Power and Potential of Generative AI*, pages 229–296. Springer.

Jörg Tiedemann and Ona de Gibert. 2023. The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327, Toronto, Canada. Association for Computational Linguistics.

Mohamed Ugas, Meredith Giuliani, and Janet Papadakos. 2024. When is good, good enough? on considerations of machine translation in patient education. *Journal of Cancer Education*, 39(5):474–476.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.

Appendix

A Hyperparameters

Model	dropout rate	learning rate grid	weight decay	batch size	epochs	optimizer
Opus-MT	0.1	1e-05, 3e-05, 5e-05, 7e-05, 1e-04, 3e-04, 4e-04, 5e-04, 2e-04, **6e-04**, 7e-04	0.01	**8**, 16, 32	3	adamw
SMaLL-100	0.1	1e-05, 3e-05, 5e-05, **7e-05**, 1e-04, 3e-04, 4e-04, 5e-04, 2e-04, 6e-04, 7e-04	0.01	**8**, 16, 32	3	adamw

Table 4: Fine-tuning hyperparameters, best in bold

B Model Performance

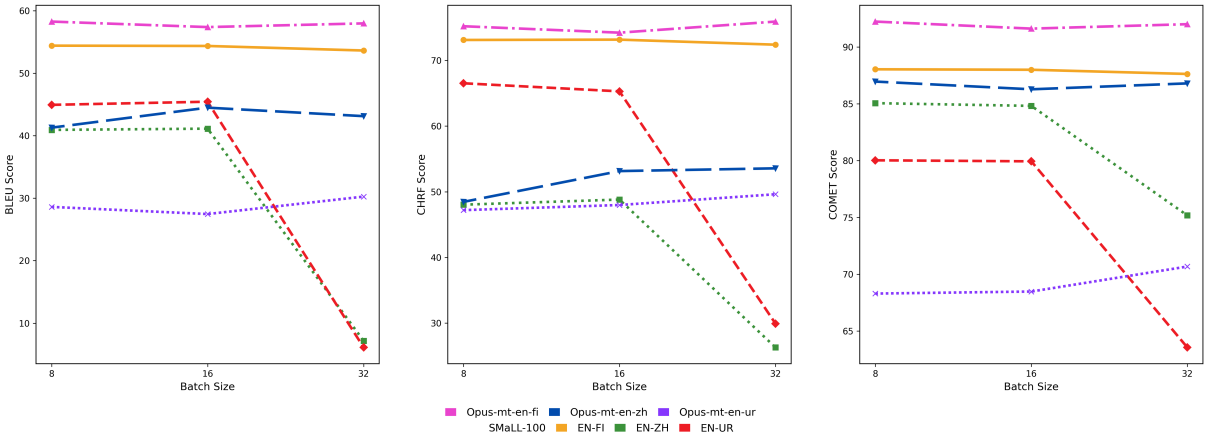

Figure 2: Evaluation metrics of the Opus-MT models and the SMaLL-100 model

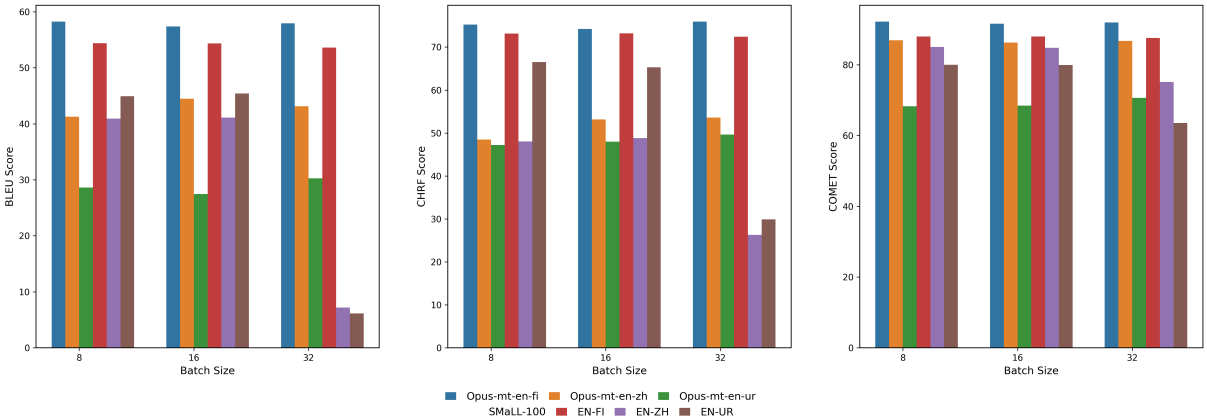

Figure 3: Comparison performance of the Opus-MT models and the SMaLL-100 model

C Automatic Evaluation

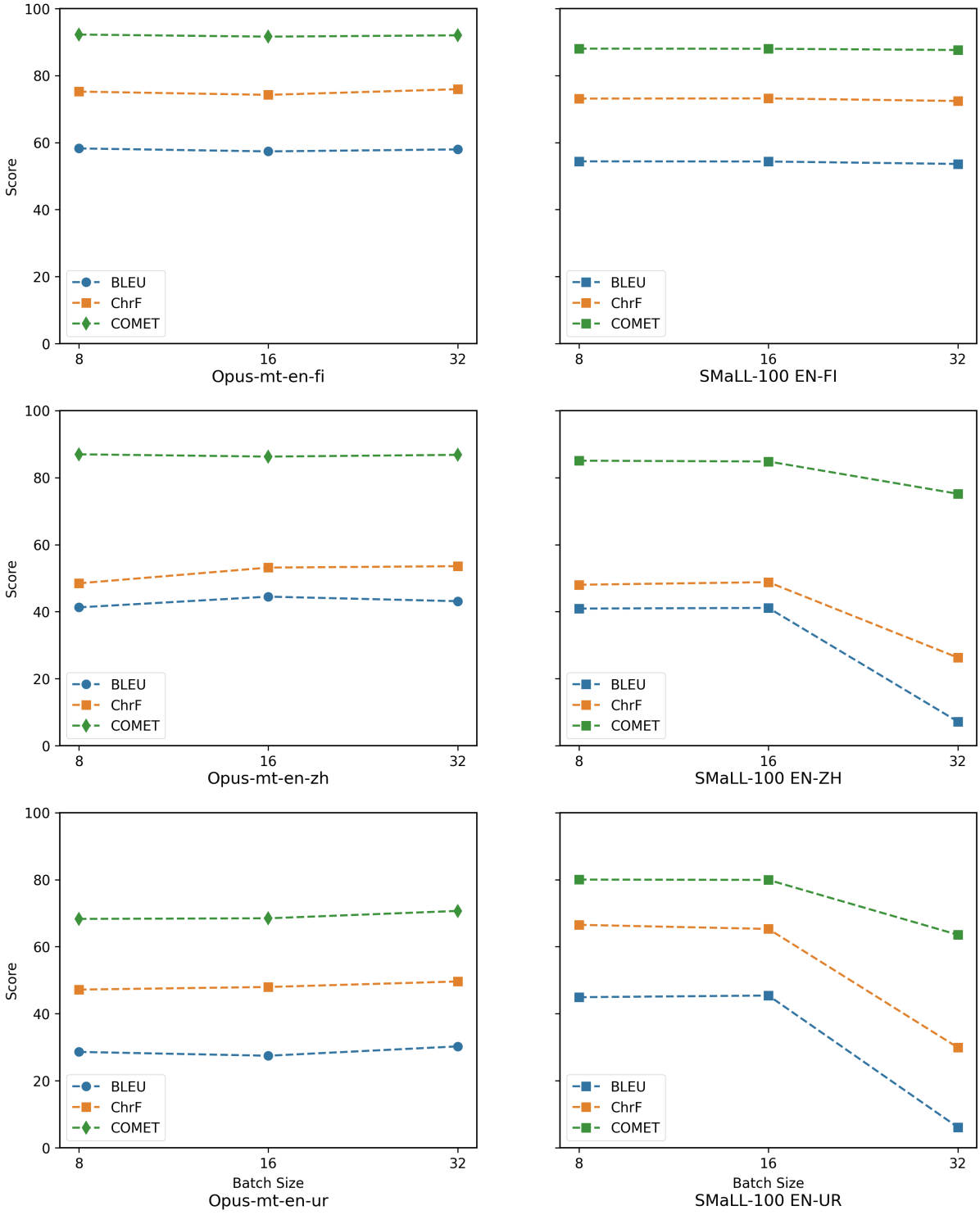

Figure 4: Performance metrics across language pairs

D Gold-Standard EN–FI Translations

Source (EN)	Reference (FI)	Target (FI)
Abdominal pain	vatsakipu	vatsakipu
Adverse effect	haitallinen vaikutus	haitallinen vaikutus
AFP levels	AFP-tasot	AFP-tasot
Alternative therapy	vaihtoehtohoito	vaihtoehtohoito
Anastrozole	anastrotsoli	anastrotsoli
Anatomy	anatomia	anatomia
Barium enema	bariumperäruiske	bariumperäruiske
Beckwith-Wiedemann	Beckwith-Wiedemann	Beckwith-Wiedemann
Blood count	verimäärä	verimäärä
Breast reconstruction	rinnan korjaus	rinnan korjaus
Bronchioloalveolar carcinoma	pienisoluinen keuhkosyöpä	pienisoluinen keuhkosyöpä
Burkitt lymphoma	Burkittin imukudossyöpä	Burkittin imukudossyöpä
Cancer staging	syövän vaiheistus	syövän vaiheistus
Chemotherapy	kemoterapia	kemoterapia
CT Colonography	paksusuolen CT-kuvantaminen	paksusuolen CT-kuvantaminen
CT imaging	CT-kuvantaminen	CT-kuvantaminen
diagnostic imaging	diagnostinen kuvantaminen	diagnostinen kuvantaminen
Digestive system	ruoansulatusjärjestelmä	ruoansulatusjärjestelmä
Epithelioid Hemangioendothelioma	solukudoskasvain	solukudoskasvain
Estrogen-only therapy	vain estrogeenihoito	vain estrogeenihoito
Febrile neutropenia	kuumeinen neutropenia	kuumeinen neutropenia
Follow-up	seuranta	seuranta
General anaesthetic	yleispuudutus	yleispuudutus
Germ cells	sukusolut	sukusolut
Hodgkin’ s lymphoma	Hodgkin-lymfooma	Hodgkin-lymfooma
High-grade dysplasia	korkea-asteinen epänormaali solukasvu	korkea-asteinen epänormaali solukasvu
Inflammatory carcinoma	tulehdusperäinen syöpä	tulehdusperäinen syöpä
Intestinal	suolisto	suolisto
key hole surgery	avainaukkoleikkaus	avainaukkoleikkaus
Kaposi sarcoma	Kaposi-sarkooma	Kaposi-sarkooma
Laparoscopic surgery	laparoskooppinen leikkaus	laparoskooppinen leikkaus
Lymph glands	imusolmukkeet	imusolmukkeet
Magnetic resonance imaging	magneettikuvaus	magneettikuvaus
Medical oncology	lääketieteellinen onkologia	lääketieteellinen onkologia
Neoadjuvant treatment	uusi hoidon tehokkuutta parantava hoito	uusi hoidon tehokkuutta parantava hoito
Nuclear medicine	isotooppilääke	isotooppilääke
Occult carcinoma	selittämätön syöpä	selittämätön syöpä
Oxaliplatin	oksaliplatiini	oksaliplatiini
Primary lymphoma	ensisijainen imukudossyöpä	ensisijainen imukudossyöpä
Palliative therapy	palliatiivinen hoito	palliatiivinen hoito
Radioactive tracer	radioaktiivinen merkkiaine	radioaktiivinen merkkiaine
Recurrent cancer	uusiutuva syöpä	uusiutuva syöpä
Sepsis pathway	sepelvaltimointerventioreitti	sepelvaltimointerventioreitti
Stage 4	vaihe 4	vaihe 4
Tissue biopsy	kudoskoepalan otto	kudoskoepalan otto
Tumor location	kasvaimen sijainti	kasvaimen sijainti
Unknown	tuntematon	tuntematon
use of statins	statiinien käyttö	statiinien käyttö
Vascular invasion	verisuonen invaasio	verisuonen invaasio
Variants	muunnokse	muunnokset
Weakness	heikkous	heikkous
Weight gain	painonnousu	painonnousu
X-ray	röntgenkuvaus	röntgenkuvaus

Table 5: A subset of accurately translated EN–FI cancer terminologies assessed with human evaluation

E Gold-Standard EN–ZH Translations

Source (EN)	Reference (ZH)	Target (ZH)
Adenopathy	腺病	腺病
Anaemia	贫血	贫血
Antibody	抗体	抗体
Anus	肛门	肛门
Artery	动脉	动脉
Assess	评估	评估
Atrophy	萎缩	萎缩
Benign	良性	良性
Cells	细胞	细胞
Colon	结肠	结肠
Dialysis	透析	透析
Diarrhoea	腹泻	腹泻
Embolism	栓塞	栓塞
Excision	切除术	切除术
Faeces	粪便	粪便
Gynaecology	妇科	妇科
Hypertension	高血压	高血压
Hysterectomy	子宫切除术	子宫切除术
Incontinence	失禁	失禁
Isotope	同位素	同位素
Laparoscopy	腹腔镜	腹腔镜
Lymph	淋巴结	淋巴
Lymphoedema	淋巴水肿	淋巴水肿
Lymphoma	淋巴瘤	淋巴瘤
Mastectomy	乳房切除术	乳房切除术
Metastasis	转移	转移
Oedema	水肿	水肿
Oncology	肿瘤学	肿瘤学
Pathology	病理学	病理学
Rectum	直肠	直肠
Recurrence	复发	复发
Relapse	复发	复发
Risk	风险	风险
Sarcoma	肉瘤	肉瘤
Screening	筛查	筛查
Side-effect	副作用	副作用
Specimen	样本	标本
Staging	分期	分期
Surgery	手术	手术
Tissue	组织	组织
Tumour	肿瘤	肿瘤
Urethra	尿道	尿道
adjuvant chemotherapy	辅助化疗	辅助化疗
allergic reaction	过敏反应	过敏反应
carcinoma in situ	原位癌	原位癌
chronic pain	慢性疼痛	慢性疼痛
clinical trial	临床试验	临床试验
digestive system	消化系统	消化系统
germ cells	生殖细胞	生殖细胞
informed consent	知情同意	知情同意
local anaesthetic	局部麻醉	局部麻醉
neoadjuvant treatment	新辅助治疗	新辅助治疗
quality of life	生活质量	生活质量
sentinel node	前哨淋巴结	前哨淋巴结
small bowel	小肠	小肠
soft tissue	软组织	软组织

Table 6: A subset of accurately translated EN–ZH cancer terminologies assessed with human evaluation

F Gold-Standard EN–UR Translations

Source (EN)	Reference (UR)	Target (UR)
Ablation Techniques	مٹانے کے طریقے	مٹانے کے طریقے
Acute hepatitis	سوزش کی جگر تیز	سوزش کی جگر تیز
alternative therapy	متبادل علاج	متبادل علاج
Anorexia	بھوک کی کمی	بھوک کی کمی
Better tolerability	بہتر برداشت (علاج)	بہتر برداشت (علاج)
Breast Self-examination	چھاتی کا خود معائنہ	چھاتی کا خود معائنہ
cancer prevention	سرطان کی روک تھام	سرطان کی روک تھام
Chemotherapy risks	کیمیائی علاج کے خطرات	کیمیائی علاج کے خطرات
chronic pain	دائمی درد	دائمی درد
Clinical trials	طبّی تجربات	طبّی تجربات
Combination chemotherapy	مجموعی کیمیائی علاج	مجموعی کیمیائی علاج
Contamination: None	آلودگی : کوئی نہیں	آلودگی : کوئی نہیں
Contralateral Disease	مخالف طرفی بیماری	مخالف طرفی بیماری
Diagnostic imaging	تشخیصی امیجنگ	تشخیصی امیجنگ
Discomfort	تکلیف	تکلیف
Dominant Geographical Areas	غالب جغرافیائی علاقے	غالب جغرافیائی علاقے
Dose/Trial Drug	آزمائشی/خوراک دوا	آزمائشی/خوراک دوا
Early pregnancy	ابتدائی حمل	ابتدائی حمل
Environmental factors	ماحولیاتی عوامل	ماحولیاتی عوامل
Excessive alcohol use	ضرورت سے زیادہ شراب کا استعمال	ضرورت سے زیادہ شراب کا استعمال
Family history	خاندانی تاریخ	خاندانی تاریخ
Follow-up	تجزیہ: پیروی	تجزیہ: پیروی
General Information About Small Cell Lung Cancer	چھوٹے سیل پھیپھڑوں کے سرطان کے بارے میں عمومی معلومات	چھوٹے سیل پھیپھڑوں کے سرطان کے بارے میں عمومی معلومات
Genetic risk factors	جینیاتی خطرے کے عوامل	جینیاتی خطرے کے عوامل
Hepatitis B	جگر کی سوزش بی	جگر کی سوزش بی
Hoarseness	آواز کا بیٹھ جانا	آواز کا بیٹھ جانا
Incidence and Mortality	واقعات اور اموات	واقعات اور اموات
Internal Validity : Fair	داخلی توثیق: معتدل	داخلی توثیق: معتدل
International Comparisons	بین الاقوامی موازنہ	بین الاقوامی موازنہ
Local radiation therapy	مقامی ریڈی ایشن علاج	مقامی ریڈی ایشن علاج
Low-birth-weight infants	کم پیدائشی وزن کے نوزائیدہ	کم پیدائشی وزن کے نوزائیدہ
Male breast cancer is rare	مردوں کے چھاتی کا سرطان نایاب ہے	مردوں کے چھاتی کا سرطان نایاب ہے
Occult NSCLC	پوشیدہ این ایس سی ایل سی کا علاج	پوشیدہ این ایس سی ایل سی کا علاج
Other risk factors	دیگر خطرے کے عوامل	دیگر خطرے کے عوامل
Overdiagnosis	ضرورت سے زیادہ تشخیص	ضرورت سے زیادہ تشخیص
Palliative therapy	تسکینی علاج	تسکینی علاج
Pathologic Classification	مرضیاتی درجہ بندی	مرضیاتی درجہ بندی
Patient Evaluation	مریض کا جائزہ	مریض کا جائزہ
Physical activity	جسمانی سرگرمی	جسمانی سرگرمی
Population-level interventions	آبادی کی سطح پر مداخلت	آبادی کی سطح پر مداخلت
Presurgical chemotherapy	سرجری سے پہلے کی کیمیائی علاج	سرجری سے پہلے کی کیمیائی علاج
Prognosis–legacy	پیش گوئی (پرانا)	پیش گوئی (پرانا)
recurrent rectal cancer	بار بار ہونے والا مستقیمی سرطان	بار بار ہونے والا مستقیمی سرطان
Screening Intervention	اسکریننگ مداخلت	اسکریننگ مداخلت
Special Populations	مخصوص آبادی	مخصوص آبادی
Stage explanation–legacy	مرحلے کی وضاحت (پرانا)	مرحلے کی وضاحت (پرانا)
Standard treatment	معیاری علاج	معیاری علاج
Study Design: Evidence obtained from large databases	مطالعہ کا ڈیزائن: بڑے ڈیٹا بیس سے حاصل شواہد	مطالعہ کا ڈیزائن: بڑے ڈیٹا بیس سے حاصل شواہد
The comparison group was not actively followed	موازنہ گروپ کی فعال نگرانی نہیں کی گئی	موازنہ گروپ کی فعال نگرانی نہیں کی گئی
The overall 5-year survival rate is 64%	کل 5 سالہ بقا کی شرح 64% ہے	کل 5 سالہ بقا کی شرح 64% ہے
To assess the efficacy of initial therapy	ابتدائی علاج کی کارکردگی کا جائزہ لینا	ابتدائی علاج کی کارکردگی کا جائزہ لینا
Tumor Characteristics	رسولی کی خصوصیات	رسولی کی خصوصیات
Weight gain	وزن میں اضافہ	وزن میں اضافہ
Who is at Risk	خطرے میں کون ہے	خطرے میں کون ہے

Table 7: A subset of accurately translated EN–UR cancer terminologies assessed with human evaluation

G Translation Errors

Error Type	Source	Reference	Target
Skipped Translations	Exemestane radiation therapist Deaths: 10,990	antineoplastinen lääke 放射治疗师 اموات: 10,990	exemestane اموات: 10,99
Grammatical Errors	GP (general practitioner) Consistency: Consistent	GP（yleislääkäri） 一致性：一致	GP(yleinen lääkäri) 一致性: 一致
Ambiguous Terms	Ablation Techniques Vancomycin Resistant Enterococcus Advanced age	ablaatiotekniikat 万古霉素耐药肠球菌 عمر رسیدگی	kudospoistotekniikat the 霉素抗性肠杆菌 دور عمر

Table 8: Some translation errors observed with human evaluation