Constituting a dataset for applying Natural Language Inference to Chinese Clinical Trials: possible approaches and challenges

Mathilde Aguiar, Ying Lai, Pierre Zweigenbaum, and Nona Naderi

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
{first.last}@universite-paris-saclay.fr

Abstract. Natural Language Inference is a Natural Language Understanding task aiming to determine the entailment relation between a given premise and hypothesis. Where we apply this task to clinical trials. We consider Clinical Trial Reports (CTR) as premises and want to determine their entailment relation with a given hypothesis. Currently, the only dataset available for this task focuses only on English. Therefore, we propose two approaches to build a new dataset using clinical trials in Chinese. The first approach consists of collecting original CTRs from Chinese clinical trial databases. The second approach uses neural machine translation models to translate the NLI4CT dataset from English to Chinese. We evaluated a sample of the our translations against human references and obtained a SacreBLEU score up to 29.11 and a BertScore of 83.62 using mBART.

Keywords: Natural Language Processing · Natural Language Inference · Clinical Trials · Multilinguality · Chinese clinical data.

1 Introduction

Natural Language Inference (NLI) is a Natural Language Understanding task that aims to determine if a hypothesis entails or contradicts a given premise. In the general domain, XNLI [2] is a multilingual large-scale dataset, with a Chinese subset, obtained by automatic translation. OCNLI [6] is a manually annotated NLI dataset built using original Chinese sources. NLI can also be applied to clinical trials. For instance, one of the applications is to enroll patients to participate in trials by using their medical profile as the hypothesis and the clinical trial's inclusion and exclusion criteria as the premise. This task has been covered in English through the dataset NLI4CT [7]. However, to the best of our knowledge, there is no dataset addressing this issue in any other languages. This could facilitate the processing of clinical trials in their original language and help build tools that could benefit medical practitioners with limited English comprehension.

Our main contribution consists of two methods to build a Chinese clinical trials corpus for Natural Language Inference. In Section 2, we present the methods we employed to build our dataset, then in Section 3 our preliminary results, in Section 4, we discuss several aspects that could be improved, and in Section 5 we conclude and give ideas for future work.

2 Methods

2.1 Collecting original Chinese clinical trials

Our first method consists of building a dataset from scratch. For premises, we want to extract clinical trial reports initially published in Chinese or both in English and Chinese. We try to extract CTRs from chictr.org.cn and chinadrugtrials.org.cn, the two official Chinese clinical trials registries. However, for both websites, each CTR needs to be downloaded one by one; on ChiCTR, the data is downloadable in an XML format, but the resulting file is in English, and on ChinaDrugTrials, each CTR needs to be downloaded separately, and the resulting .doc file has broken the original CTR's formatting. In addition to data acquisition issues, we also face annotation issues. Indeed, to build an NLI dataset, we need annotators to produce hypotheses, and in the case of clinical trials, we need expert annotators with clinical knowledge. Since these annotators are costly and hard to recruit, we came up with a second solution based on automatic translation.

2.2 Automatic translation of English NLI4CT

This approach uses Neural Machine Translation (NMT) models to translate the English NLI4CT dataset. This approach allows us to translate a large quantity of text without the cost of manual translation. Moreover, similar NMT approaches [1,5] have obtained reliable translations. We tried different kinds of pre-trained models to translate the dataset from English to Chinese: (I) OPUS-MT-zh-en [10], a Chinese-English NMT model pretrained on general domain data, (II) mBART [9] and (III) M2M-100 [3], 2 multilingual NMT models also pretrained on general domain data, and (IV) Taiyi [8] a bilingual Chinese-English Large Language Model (LLM) instruction-tuned on a set of biomedical tasks, including biomedical English-Chinese translation. As a result, we obtained a dataset with 1,700 training instances, 200 instances for development and 500 for testing, as the original English version.

3 Results

To evaluate the quality of the translations, we randomly sampled 50 sentences from the original English NLI4CT and asked one of our authors, a native Chinese speaker, to provide a translation reference for each of these sentences. Using these references, we compute the BertScore, the SacreBLEU, ChrF, and TER scores by comparing the human references with the model's translations.

mBART obtained the best results among all the baselines, followed by M2M-100. Despite being the larger model and instruction-tuned over biomedical tasks,

Table 1. Automatic metrics of the obtained translations against our human references. The evaluation sample consists of 50 sentences.

Model	SacreBLEU (↑)	BertScore (↑)	ChrF (↑)	TER (↓)
Opus-mt-zh-en	24.21	80.53	26.53	**96.61**
mbart-large-50-one-to-many-mmt	**29.11**	**83.62**	**31.30**	107.63
m2m100_1.2B	26.53	82.27	28.64	110.17
Taiyi-LLM	17.41	75.54	22.79	128.81

Taiyi obtained the worst results. OPUS, being around 100 times smaller than Taiyi, still performs better. For our use case, models specifically built for translation perform better than our LLM baseline.

4 Discussion

Although a solution based on NMT is quick and easy to set up, it has several weaknesses. As shown by the automatic metrics, the translation quality is still to be improved. Medical text is challenging to translate due to numerous domain-specific terms, abbreviations, and special syntax. Moreover, it is possible that by translating the instances, the semantic changes and the original label no longer correspond to the translated version of the hypothesis and statement [4]. In addition, it is necessary to expand the size of the human references for translation to evaluate the obtained translations more accurately.

The original NLI4CT is solely based on Breast Cancer CTRs, consequently leading our dataset to only cover Breast Cancer CTRs. Having CTRs related to more diseases should be considered to broaden our task.

5 Conclusion and Future Work

In this paper, we presented our approaches to build a Chinese version of NLI4CT, a dataset for applying Natural Language Inference to clinical trials. We first try to collect clinical trial reports in Chinese from official Chinese clinical registries. This approach yielding several challenges, we adopt a second approach consisting of automatically translating the English dataset NLI4CT. With this approach, we obtained a dataset of 2,400 instances. We also evaluated the quality of the resulting translation by evaluating a sample of instances against human references. Our work is still in progress, and we wish to expand our human references for evaluation and explore other techniques to obtain better translation quality, such as using entity linking for a better translation of medical terms, or by finetuning our translation models on biomedical Chinese-English parallel corpora for translation. We also want to finetune Masked Language Models on our resulting dataset to evaluate its difficulty compared to the English version and check for potential biases. The code is available on our GitHub[1].

[1] https://github.com/CTInfer

Acknowledgments. This work benefited from the GPUs provided by Lab-IA, an institution member of Université Paris-Saclay. This work was also supported through the CNRS grant 80|PRIME.

References

1. Aggarwal, D., Gupta, V., Kunchukuttan, A.: IndicXNLI: Evaluating multilingual inference for Indian languages. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 10994–11006. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.755, https://aclanthology.org/2022.emnlp-main.755
2. Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2018)
3. Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., Joulin, A.: Beyond english-centric multilingual machine translation. J. Mach. Learn. Res. **22**(1) (jan 2021)
4. Ham, J., Choe, Y.J., Park, K., Choi, I., Soh, H.: KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 422–430. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.39, https://aclanthology.org/2020.findings-emnlp.39
5. Heredia, M., Etxaniz, J., Zulaika, M., Saralegi, X., Barnes, J., Soroa, A.: XNLIeu: a dataset for cross-lingual NLI in Basque. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 4177–4188. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.234, https://aclanthology.org/2024.naacl-long.234
6. Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., Moss, L.: OCNLI: Original Chinese Natural Language Inference. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 3512–3526. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.314, https://aclanthology.org/2020.findings-emnlp.314
7. Jullien, M., Valentino, M., Frost, H., O'Regan, P., Landers, D., Freitas, A.: NLI4CT: Multi-evidence natural language inference for clinical trial reports. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 16745–16764. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.1041, https://aclanthology.org/2023.emnlp-main.1041
8. Luo, L., Ning, J., Zhao, Y., Wang, Z., Ding, Z., Chen, P., Fu, W., Han, Q., Xu, G., Qiu, Y., Pan, D., Li, J., Li, H., Feng, W., Tu, S., Liu, Y., Yang, Z., Wang, J., Sun, Y., Lin, H.: Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. Journal of the American Medical Informatics Association **31**(9),

1865–1874 (02 2024). https://doi.org/10.1093/jamia/ocae037, https://doi.org/10.1093/jamia/ocae037

9. Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A.: Multilingual translation with extensible multilingual pretraining and finetuning. ArXiv **abs/2008.00401** (2020), https://api.semanticscholar.org/CorpusID:220936592
10. Tiedemann, J., Thottingal, S.: OPUS-MT — Building open translation services for the World. In: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT). Lisbon, Portugal (2020)