Experience Report: Implementing Machine Translation in a Regulated Industry

Per Fallgren, Marco Zocca, David Buffoni
Mölnlycke Health Care
{first.last}@molnlycke.com

Abstract

This paper presents lessons learned from implementing Machine Translation systems in the context of a global medical technology company. We describe system challenges, legal and security considerations, and the critical role of human-in-the-loop validation for quality assurance and responsible deployment. Furthermore, based on an experiment involving over 11,000 ranked translations, we report reviewer preferences for outputs from small and large language models under various prompting configurations, using a domain-specific dataset spanning five language pairs.

1 Introduction

Companies with a global presence typically invest substantial resources in translating content into the various languages required across their global markets. This work is often outsourced to third-party providers, who are entrusted with both data security and the delivery of high-quality translations. However, recent breakthroughs in neural machine translation (NMT) and large language models (LLMs) have significantly improved the quality and accessibility of Machine Translation (MT) tools, making them a plausible alternative to traditional services.

High accuracy and compliance to standard terminology is a critical requirement of a translation system to be employed in the medical industry; any false or misleading claims give rise to regulatory scrutiny or, worse, adverse medical outcomes. Implementing automated translation in our context effectively aims at replacing the judgement of few select domain experts with an AI system of comparable proficiency, which is a substantial undertaking.

The challenge is compounded by a lack of domain-specific evaluation data across all languages and tones of interest (making this effectively a "low-resource" translation setting), which provided the starting point for our experimentation.

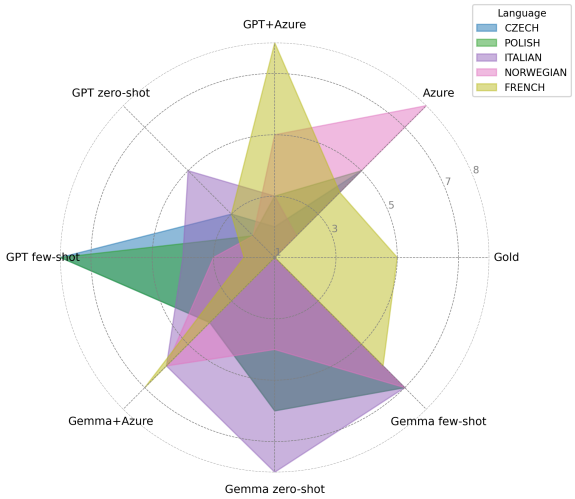


Figure 1: Aggregate human validator rankings of 7 MT methods (1 is best) over our dataset of 5 language pairs. No single method of the ones we evaluated dominates reviewer preference across all language pairs, the "Gold" control ranks poorly in one language and the lower ranks have high variance. Refer to §3 for details on the experimental protocol and results. In this figure we aggregate rankings using the Borda algorithm, as implemented in Pereira and Pettit (2025).

This paper makes the following contributions:

- We provide an account of the challenges encountered and lessons learned from evaluating in-house new MT tools and workflows, and
- we present the results of an experiment we carried out to evaluate and compare different MT tools, offering insights into selecting the most suitable method for specific translation tasks.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1667–1673

2 Lessons learned

Mölnlycke Health Care, a global MedTech company with products distributed in over 100 countries, provided a representative environment for exploring the complexities of multilingual AI deployment. Organizations of comparable size and international reach are likely to encounter similar challenges. The lessons learned and documented herein can hopefully support the implementation of machine translation systems in similarly complex operational settings.

2.1 Selection of initial use cases

At Mölnlycke Health Care, translation needs span the entire organization, from public-facing web content to highly regulated documents such as instructions for use of medical products. In light of this diversity, it is advisable to begin with low-risk, low-complexity use cases. This approach minimizes the potential impact of early stage issues, facilitates leadership approval, and enables the gradual development of a robust MT strategy that can later be extended to more complex and sensitive content.

2.2 Cross-functional involvement

Building a fully in-house translation workflow requires coordinated involvement across multiple functions within the organization. While each company has its own structure and strategy, a key recommendation is to identify and engage all relevant stakeholders as early as possible in the development process.

This may include technical teams, not only those working with AI, but also experts in IT security and architecture. Legal and data privacy teams should also be consulted to ensure compliance. Internal validators with both language proficiency and domain expertise should be identified early. Supporting this process may require the use of an annotation platform. If such a platform is developed in-house, the involvement of UX designers and front-end developers can be particularly valuable.

2.3 Leveraging existing translation data

Established organizations with a global presence possess large volumes of previously translated material. We recommend making a deliberate effort to collect and utilize this data, as it can significantly enhance the performance and domain alignment of translation tools.

Many off-the-shelf MT systems support fine-tuning, allowing organizations to adapt models using their own data to better reflect company-specific tone of voice and terminology. In addition, translation memories can be extracted from existing content, enabling consistent reuse of validated translations.

If third-party translation services have been used, it is likely that translation memories and terminology databases have already been created and stored. These are often company-owned resources that can be repurposed internally. Using these resources can speed up development and improve translation quality.

2.4 Validation

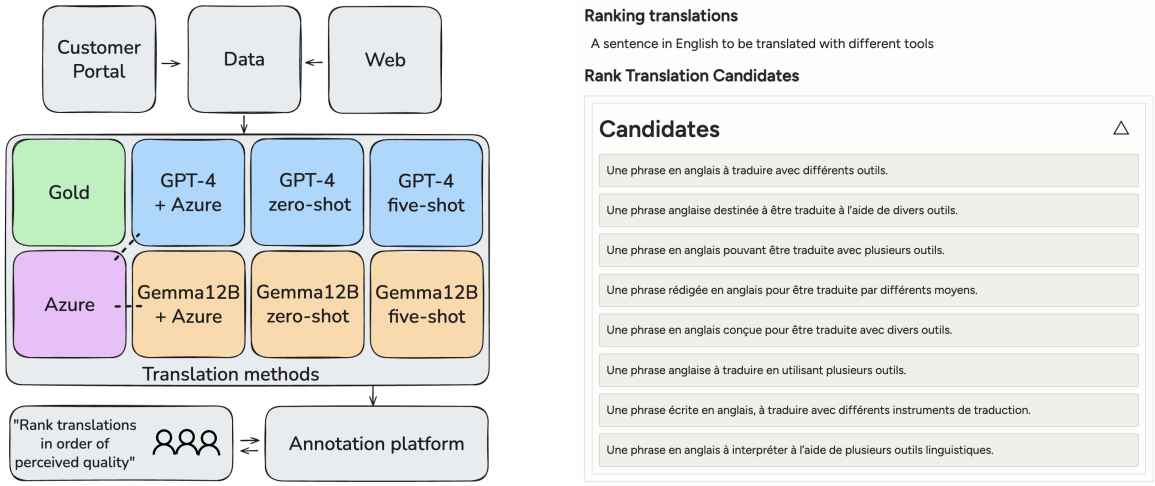
In the context of this experiment, "validation" means evaluating that an automatically translated text is acceptable with respect to meaning and fluency. This task requires groups of proficient speakers of the relevant language pairs, and who are also conversant in the relevant technical jargon and shorthand; finding enough reviewers that meet these criteria proved to be challenging in our setting. In some cases, commercial terminology did not have a single or well-defined translation in the target language, causing uncertainty among the reviewers.

A user-friendly validation platform not only reduces cognitive load for validators but can also integrate translation memory matches; these can be presented alongside the content, and dynamic rephrasing suggestions or synonym recommendations could further enhance the user experience.

In typical translation annotation workflows, human annotators validate translations by editing and approving them. Instead, the experiment presented in this paper initially asks annotators to rank translations produced by various systems. The motivation is to identify the most suitable translation tool for a given use case, thereby potentially reducing long-term effort. Once the optimal system is selected, post-editing can still be applied if necessary, but now using outputs from the most appropriate tool.

2.5 Data governance

Implementing a translation validation software platform in a regulated industry is made challenging by competing requirements of technology integration, data security and compliance etc. In addition



(a) (b)

Figure 2: (a) Visualization of the data flow, translation types, and annotation process used in the experiment. (b) Screenshot of an annotation task in Label Studio, illustrating example data for English-to-French translation. Each box represents one translation generated by one of the eight evaluated approaches. The boxes are drag-and-droppable, and the annotator's task is to rank the translations from best to worst based on perceived quality.

to the confidentiality of intellectual property and processes, one of the requirements of our industry is for all decisions that might end up in a medical device (including data annotations and documentation) to be traceable to their author and date. To meet these criteria we chose a data annotation platform that interfaces directly with our cloud object storage (i.e. not requiring any processing third parties), and that generates a "paper trail" of all authoring events.

3 Experiment

We evaluate the output of three MT models (described in §3.2), by collecting quality ranking judgements from a panel of human annotators. In each translation reviewing task (shown in Figure 2), the annotators were shown 8 translations in randomized order, of which 7 were produced by MT methods and one being the reference translation (pre-approved by company experts), which we refer to as "Gold". Figure 2 shows a visual summary of the experiment setup.

3.1 Data

We assembled a dataset of translation sentence pairs from English to five target languages (Czech, French, Norwegian, Italian, Polish), starting from 100 reference sentence pairs taken from two internal sources we describe in the following.

3.1.1 Customer Portal

The Mölnlycke Portal is a digital platform designed to support healthcare professionals in managing customized procedure trays. It is deployed across multiple markets, necessitating multilingual support. The language in these sentences includes terminology related to medical technology such as description and operation of Mölnlycke Health Care products.

For this study, 50 sentences were randomly extracted from the Portal content, each accompanied by an existing validated human translation. Very short and very long sentences were excluded.

3.1.2 Web

The part of the dataset we refer to as Web Data was extracted from previously translated and validated content that is publicly accessible via the Mölnlycke Health Care website. This content has been localised to various markets and the domain is specific to Healthcare and promoting Mölnlycke Health Care products. For this study, we clustered the raw text using BERTopic (Grootendorst, 2022) and uniformly sampled 50 sentences from the resulting clusters to ensure uniformity across the Mölnlycke Health Care business applications. These sentences were translated into various languages and the translation was validated by domain experts. Very short and very long sentences were removed similarly as in the Customer Portal Data.

Prompt Type	Instruction
System Message	`You are an expert in English and {target_language}.` `Please provide a high-quality translation of the provided text from English to {target_language}. Only generate the translated text. No additional text or explanation needed.`
Zero-shot	`Translate the following text into {target_language}:` `{source_sample}`

Few-shot	`Translate the following text into {target_language}: {source_shot1}`
	`{validated_shot_translation1}`
	`...`
	`Translate the following text into {target_language}: {source_shot5}`
	`{validated_shot_translation5}`
	
	`Translate the following text into {target_language}: {source_sample}`

Refinement	`The following text has been translated into {target_language}.`
	`Refine the translation to improve accuracy, fluency, and faithfulness to the original English. Use natural language, preserve the original meaning, and fix any errors or awkward phrasing.`
	
	`Existing translation: {azure_translation}`
	`English master: {source_sample}`

Table 1: Prompt types and corresponding instructions used for the three configurations of Gemma-12B and GPT-4.

3.1.3 Data generation

The two data sources, each containing 50 examples, served as the dataset for generating translations. Each English example was translated using the seven approaches described previously (see Section 3.2), including the human-validated reference we then get eight translations. This resulted in a total of 4,000 translations (100 data samples × 8 translation approaches × 5 target languages) for annotation.

For the five-shot configurations, five additional example pairs per language, comprising both source (English) and target translations, were extracted from the same data sources and incorporated into the language model prompts according to Table 1.

3.2 Machine translation methods

We evaluated seven automated translation methods, representing three commonly available categories in contemporary MT: traditional neural machine translation (NMT), large language models (LLMs), and smaller language models (SLMs).

The first approach is *Microsoft Azure Translator*, a widely adopted NMT solution. The second employed *GPT-4*[1] (OpenAI et al., 2024) via the Azure OpenAI Service, a standard LLM-based method. The third involved a quantized *Gemma 12B* model (Mesnard et al., 2024), accessed through Ollama[2], a lightweight SLM-based approach.

Each language model was tested using three distinct prompting strategies:

1. **Zero-shot translation** – a direct instruction to translate from English to the target language without additional context.

2. **Few-shot translation** – the model was primed with five example translations prior to generating new ones.

3. **Refinement-based translation** – the model received the Azure-generated translation and was prompted to improve its quality.

A summary of the language model configurations is provided in Table 1.

[1] GPT-4o, model version 2024-11-20

[2] gemma3:12b-it-qat, downloaded 2025-05-20 from https://ollama.com/library/gemma3:12b

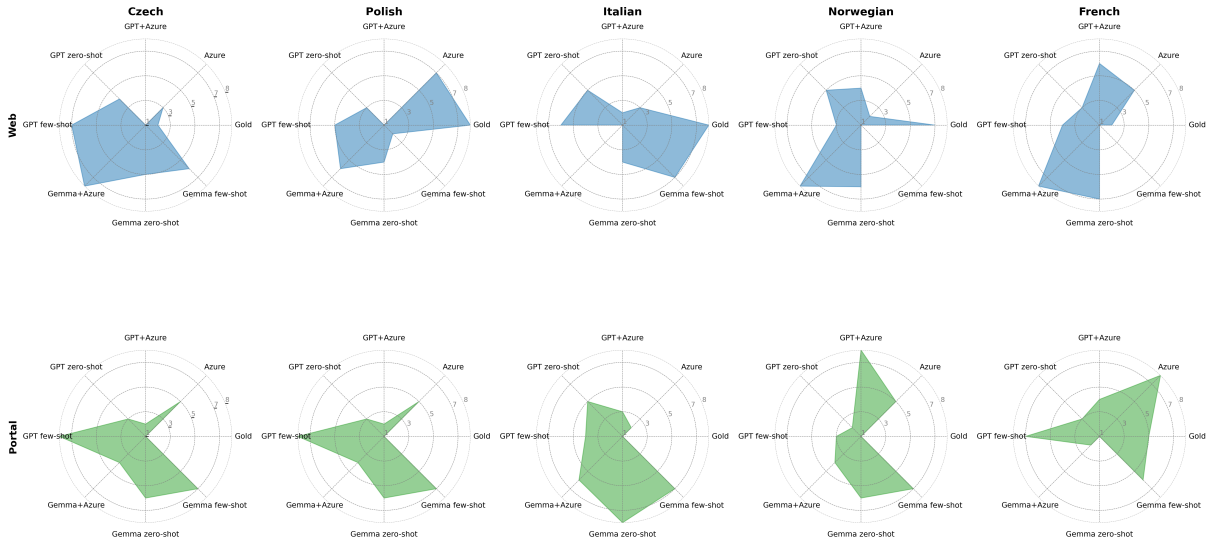


Figure 3: Aggregate preference rankings for our 5 target languages (1 is best, Borda method) controlling for the data source of the original text.

Language	A_{min}	A_{max}	A_{avg}	RBO (p = 0.75)
Czech	4	5	4.01	0.54
French	1	3	2.06	0.51
Italian	2	4	2.88	0.58
Norwegian	1	2	1.99	0.54
Polish	2	3	2.98	0.51

Table 2: Summary statistics of the labeled dataset. A denotes the number of annotations per task (one task being a ranking of 8 translations), RBO the Rank Biased Overlap score (Webber et al., 2010).

3.3 Data annotation

We used Label Studio, an open-source data labeling platform, to manage the annotation process. The experiment utilized the platform's ranker tag, which enabled annotators to drag and drop translation candidates into a ranked order based on perceived quality (see Figure 2).

Sixteen native speakers participated in the annotation process, covering the following languages: French (3), Norwegian (2), Czech (5), Italian (3), and Polish (3). A total of 11,136 translations were ranked (based on 1392 produced annotations for 8 translation types), and Table 2 shows some summary statistics of the resulting labeled dataset.

3.4 Results and analysis

Our translations ranking dataset as described in §3.3 is the starting point for our analysis. We first aggregate rankings using the Borda algorithm (Wang et al., 2024), as implemented in Pereira and Pettit (2025).

The per-language RBO score we report in Table 2 shows substantial inter-annotator agreement in the top positions of the produced rankings.

Figure 3 shows a summary of our current results: the most evident result is that no single MT method of the ones we evaluated is universally preferable across all language pairs.

For all languages except French, "Gold" is the preferred alternative (compared to a random choice, 1/8), from 27 to 42% of the time, this is also reflected for each validator who selected "Gold" as their favorite alternative. For French, this preference is less than the uniform probability (11%) and no annotator considered "Gold" as their favorite.

Looking at the variability of the reviewer preferences for the "Gold" translation with respect to our two data sources (Figure 3), we notice the effect of reviewer domain knowledge; "Portal" data appears as simple to translate as the rankings clearly show the reviewers prefer the reference translation, whereas healthcare-specific terminology used in the "Web" source text gives rise to higher reviewer uncertainty (annotators did not rank "Gold" as first choice 55 to 65% of the time). We suspect that our experimental protocol may reinforce this effect because source sentences are short implying small variations among translations, translation alternatives may be exactly the same, and the displayed order on the annotation platform is randomized. We report an example in French in Table 3 where we can see 3 MT alternatives exactly like the "Gold" one. In this particular case, the 3 French validators produced: - 1 kept the displayed order as it was presented on the platform

Source	*What is the most effective way to apply the XXX solution to the skin?*
Gemma few-shot	*Quelle est la manière la plus efficace d'appliquer la solution XXX sur la peau ?*
Gemma 1-shot	*Quelle est la méthode la plus efficace pour appliquer la solution XXX sur la peau ?*
Azure	*Quelle est la façon la plus efficace d'appliquer la solution XXX sur la peau ?*
Copilot (Azure)	*Quelle est la méthode la plus efficace pour appliquer la solution XXX sur la peau ?*
Gold	*Quelle est la manière la plus efficace d'appliquer la solution XXX sur la peau ?*
Copilot few-shot	*Quelle est la manière la plus efficace d'appliquer la solution XXX sur la peau ?*
Copilot 1-shot	*Quelle est la manière la plus efficace d'appliquer la solution XXX sur la peau ?*
Gemma (Azure)	*Quelle est la méthode la plus efficace pour appliquer la solution XXX sur la peau ?*

Table 3: Example of an ambiguous translation ranking task, from the "Web" fraction of the English-French sentence pair dataset. For short sentences, some translation methods give identical results.

but this behavior appeared rarely in our experiment (less than 5% of time among languages) - 1 clustered the "Gold", "Gemma Few-shots", "Copilot Few-shots" and "Copilot Zero-shot" in the first ranks but "Gold" appears in the top-4 for this particular example. In our experiments, except for French, Gold has the highest probability to appear in the top-3 compared to other systems. - 1 chose another type of alternatives, the one provided by "Gemma Zero-shot" where the term "méthode" was preferred instead of "manière" in the translation. In that situation, there is a subtle difference between the two terms such as in English between "method" and "way". This last behavior seems to explain why "Gold" is not the most preferred option among systems. We think this is due to the competing effect of validators' expertise and the high quality of recent AI translation models for non-domain specific data (e.g. the Web fraction of the dataset).

4 Discussion

Our current results give us several directions for further exploration. On the experimental side, it is striking that the "Gold" (pre-validated) set of translations is not ranked as favorite in at least one language setting (e.g. French). We think this is due to the competing effect of validators' expertise and the high quality of recent AI translation models for non-domain specific data (e.g. the Web fraction of the dataset). Controlling for keywords and domain-specific language, as well as reviewers' expertise and prior exposure to the respective data sources, could clarify why AI-generated translations are preferable to the "Gold" fraction in some cases.

Given the high cost of human validation, one option for the future could be to assess how well an LLM can rank translations; if its rankings correlate strongly with human judgments, it could serve as a viable and more cost-effective alternative, but could also introduce harmful biases.

5 Related work

Domain-specific language data Constructing meaningful translation datasets in specific domains is a common challenge in industry, and in a validated MT system the human annotators are usually augmented by language models at various stages (see e.g. Kang et al. (2025)).

Domain adaptation of language models Even though modern foundation language models display impressive fluency on a range of tasks, their aptitude on domain-specific text is far from granted. This is an instance of Domain Adaptation (DA) (Marashian et al., 2025), and both prompting (Peng et al., 2023) and fine-tuning (Xu et al., 2019) on tasks of interest have been proven to be effective for improving output quality.

Translation post-editing It is very complex to edit and approve translations produced by machine translation systems (Pérez, 2024). Misalignments often arise between industry and translators, leading to calls for new guidelines better suited to this work.

6 Conclusion

In this paper we have presented some challenges and lessons learned while implementing AI-based language translation at scale in a regulated industry. Adopting an annotation tool that provides a comfortable UX while complying with regulatory requirements was a useful starting point, but uniform data coverage and quality proved to be challenging. Regarding AI model selection, we do not observe a single one that ranks best among the set we considered, suggesting the need to involve different translation methods depending on language

and domain.

A central component of this study is an experiment involving 4,000 translations across five languages, evaluated by native-speaking internal reviewers. The experiment compares outputs from human translators, large language models (LLMs), Azure Translator, and hybrid approaches using a dedicated annotation platform. Our findings highlight not only the comparative performance of these tools but also the significance of user experience (UX) in annotation workflows. By sharing our methodology and insights, we aim to provide a practical framework for other organizations to assess and optimize MT tools tailored to their specific needs.

Limitations

An uneven coverage of reviewers for our language pairs creates rankings of varying quality, which may bias the results.

Our study considered a fairly representative group of AI language translation models, but this is by no means exhaustive; in particular, we did not have the resources to evaluate translation models that are specific to language pairs, e.g. OPUS-MT (Tiedemann and Thottingal, 2020). The model prompts we use are static, i.e. do not introduce context-specific information but only translation patterns.

Since the translations were sourced from previously validated data, it cannot be guaranteed that the validators had no prior exposure to the material.

Parts of the dataset used in this study is internal to our organization, which prevent us from sharing it externally.

Acknowledgments

We thank the volunteers who generously contributed their time and language expertise to help annotate translations for this study.

References

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *Preprint*, arXiv:2203.05794.

Junghoon Kang, Keunjoo Tak, Joungsu Choi, Myunghyun Kim, Junyoung Jang, and Youjin Kang. 2025. DaCoM: Strategies to construct domain-specific low-resource language machine translation dataset. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 612–624, Abu Dhabi, UAE. Association for Computational Linguistics.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, and 88 others. 2024. Gemma: Open models based on Gemini research and technology. *Preprint*, arXiv:2403.08295.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Valdecy Pereira and Brigham Pettit. 2025. PyRankMCDA - rank aggregation methods for MCDA problems. GitHub.

Celia Rico Pérez. 2024. Re-thinking machine translation post-editing guidelines. *The Journal of Specialised Translation*, (41):26–47.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Siyi Wang, Qi Deng, Shiwei Feng, Hong Zhang, and Chao Liang. 2024. A survey on rank aggregation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Jitao Xu, Josep Crego, and Jean Senellart. 2019. Lexical micro-adaptation for neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.