UniOR-MT: An Agentic Pipeline for Terminology Translation

Antonio Castaldo[1,2], Giulia Speranza[1], Maria Carmen Staiano[3], Petra Giommarelli[1,2], Johanna Monti[1], Maria Pia Di Buono[1]

[1]University of Naples "L'Orientale" [2]University of Pisa [3]University of Macerata

antonio.castaldo@phd.unipi.it, gsperanza@unior.it, m.staiano@unimc.it, petra.giommarelli@phd.unipi.it, jmonti@unior.it, mpdibuono@unior.it,

Abstract

We present UniOR-MT, an agentic terminology-aware machine translation pipeline developed for the WMT 2026 Terminology Translation Shared Task. We compare five progressively enriched translation paths, each adding one component to isolate its individual contribution. We additionally submit a mixed system that selects, among the five candidate paths, the best candidate using a combined measure of terminology accuracy and document-level quality estimation. Evaluation results show that incorporating a bilingual glossary provides the most consistent benefit, while additional post-editing steps yield limited improvements, and can even degrade translation quality for Spanish–Basque in the official evaluation. The final system, which includes iterative QA feedback and a stronger fallback model, achieves the highest docCOMET scores in our internal evaluation.

1 Introduction

The translation of specialised terminology has been studied for decades, dating back to early work in the 1960s (Dunne, 2013). Regardless of whether the users are domain experts or professional translators, the primary objective is the accurate transfer of specialized knowledge (Pavel and Nolet, 2001). Consequently, translation-oriented terminology work focuses on ensuring consistency and accuracy (Prieto Ramos, 2021), enabling specialised information to be conveyed clearly to the end user. Meeting these requirements remains a major challenge for machine translation (MT) systems. Despite substantial advances in MT research, terminology-heavy domains remain far from being fully resolved. While general-purpose systems have achieved near-human performance for many high-resource language pairs, high-stakes domains such as science, engineering, medicine, and finance continue to pose significant challenges. Modern MT systems often demonstrate strong sentence-level performance, but ensuring terminological consistency across documents remains an open research problem (Semenov et al., 2025).

These challenges provide the motivation behind the fourth edition of the WMT 2026 Terminology Translation Shared Task (Charkiewicz et al., 2026), which focuses on document-level translation performed under explicit terminology constraints. Extending the scope of previous editions, this year's task broadens the evaluation to include lower-resourced, morphologically rich languages, offering a more challenging benchmark for terminology-aware machine translation systems. The WMT 2026 Terminology Task includes two tracks:

- **Track 1:** Document-Level Translation with Explicit Dictionary covers Spanish–Basque in the Engineering and Technology domain, as well as English–Polish in the Engineering and Technology and Medicine domains.
- **Track 2**: Document-Level Translation with Sample Bitexts covers Traditional Chinese–English in the Finance domain.

We focus on Track 1, which targets document-level translation with an explicit terminology dictionary. For each language pair, the input is provided as a JSON file of document-length source chunks (~2,000 words each), accompanied by a corpus-level bilingual dictionary in JSON format, mapping source terms to their target-language equivalents.

Systems are evaluated under three modes of terminology translation: i) no terminology (plain document translation, with no dictionary supplied), ii) proper terminology (translation guided by the correct domain dictionary), and iii) random terminology (translation guided by a control dictionary of the same format and size as proper, but populated with words sampled arbitrarily from the input

Proceedings of the Eleventh Conference on Machine Translation, pages 1779–1790
October 28-29, 2026 ©2026 Association for Computational Linguistics

rather than domain terms). This three-way setup isolates the causal contribution of accurate terminology from the effect of simply providing any additional lexical input.

For Track 1, we propose a modular translation pipeline made up of a sequence of progressively enriched systems. Starting from a zero-shot translation baseline (ZS), the pipeline incrementally introduces bilingual terminology constraints (ZS+T), Post-editing (PE), Rule-guided Post-editing (Rule PE), and iterative Quality Assurance (QA). By adding a single module at each stage, the proposed architecture enables a controlled ablation of the contribution of each terminology-aware intervention, as presented in Table 1.

Internal analysis shows that additional stages can optimize the system on specific automatic metrics, with terminology injection (ZS+T) providing the most benefit. The official blind evaluation shows that this increased complexity does not necessarily translate into better performance, with ZS+T achieving the strongest performance among our submitted systems. Terminology injection emerges as the most robust strategy across language pairs, while additional post-editing produces small improvements for English–Polish but substantial degradation for Spanish–Basque, resulting in worse overall performance.

2 Related Work

Research on terminology-aware machine translation has been approached with a variety of methods aimed at improving the consistent use of domain-specific terminology. These methods are traditionally grouped into two strategies that differ in how strongly terminological choices are imposed on the model: hard and soft terminology constraints. Hard constraints intervene directly in the decoding process of the model to ensure that a given set of terms appear in the generated translation. These methods have been shown to achieve high terminological accuracy (Huang et al., 2024), but they often negatively affect overall translation quality and fluency (Hasler et al., 2018; Zhang et al., 2021). For this reason, they are often followed by a refinement stage, as seen in Wu et al. (2025). A more flexible alternative is offered by soft constraints, where terminology is provided as guidance during the generation process. In this methodology, bilingual glossaries and dictionaries are incorporated into the model's input, allowing the model to be guided by the given terminology without modifying the generation process itself (Briva-Iglesias and Ferre-Fernández, 2026; Hauhio and Friberg, 2024; Moslem et al., 2023).

Recent work has increasingly combined terminology-aware translation with refinement stages. Translate–then–refine approaches employ constrained re-decoding or LLM-based post-editing to improve terminology adherence while preserving translation quality (Bogoychev and Chen, 2023; Jaswal, 2025).

A paradigm that has become increasingly popular in MT is the use of collaborative agentic pipelines. Depending on the workflow design, these systems employ different collaboration strategies, including reflection-based refinement, where an LLM iteratively critiques and improves its own translation (Ng, 2025); role-based collaboration, in which specialised agents such as translators, reviewers, proofreaders, or domain experts sequentially refine translations (Wu et al., 2025; Sin et al., 2025; Briva-Iglesias, 2025); and consensus-based architectures, where multiple models or agents generate alternative translations that are subsequently reconciled by an arbitration mechanism (Staiano et al., 2026). These approaches use specialised agents to improve translation quality through iterative refinement, collaboration, or agreement before producing the final output.

A common design pattern in these systems is the generation of multiple translation candidates followed by a selection stage that balances overall translation quality against terminology accuracy, rather than optimising a single objective (Guttmann et al., 2025; Grubišić and Korencic, 2025). However, such selection is not without risks: a selector driven by a single utility metric may prefer candidates that omit the required terms, a form of metric interference (Pombal et al., 2025). These systems generate their candidates in parallel and do not isolate which stage of terminology intervention drives the trade-off between adherence and quality.

Our work addresses this limitation by adopting an incremental pipeline in which terminology-aware components are introduced progressively. This design enables a controlled evaluation of each intervention, allowing us to isolate its individual contribution to terminological accuracy and translation quality.

3 Methodology

We developed a modular MT pipeline, composed of five progressively enriched paths and relying primarily on `gpt-4o-mini` as the underlying language model. For each source segment, the pipeline generates five translations from five different paths, each corresponding to the system configurations described in Section 3.1. Each system builds directly upon its predecessor by the addition of a single component, allowing the isolated effect of each intervention to be assessed while holding all other factors constant. Using a relatively small model, the proposed architecture allows us to investigate whether structured terminology-aware interventions can achieve strong performance without requiring larger language models throughout the pipeline. Figure 1 provides an overview of the final agentic pipeline.

3.1 System Configurations

Path 1: ZS. The baseline system consists of a single language model acting as a translation agent (GPT-4o-mini), prompted to produce a direct translation of the source text with no auxiliary resources. This condition serves as the lower-bound reference against which the contribution of subsequent components is measured.

Path 2: ZS+T. Path 2 extends Path 1 by supplying the translation agent with the domain-specific bilingual dictionary provided by the workshop organisers. The dictionary is injected into the prompt as terminological constraints, with the aim of improving lexical accuracy and terminological consistency without altering the underlying translation process.

Path 3: PE. Path 3 introduces a second agent, a post-editor, which receives the output of Path 2 (translation agent + dictionary) and operates under a general instruction to improve the fluency, accuracy, and overall quality of the translation, without being given explicit editing criteria.

Path 4: Rule PE. Path 4 retains the architecture of Path 3 but replaces the generic post-editing instruction with an explicit set of post-editing rules (see Appendix A). These rules were designed by drawing on established principles from terminology theory and translation studies (Fernández-Parra, 2025; Bowker, 2015; Cabré, 2010; Scarpa, 2001). In particular, we selected rules to instruct the model to verify the consistent use of terms throughout the translation (ISO 11669:2024, 2024; ISO 17100:2015, 2015), ensure compliance with terminological databases (Risku, 2006), evaluate whether the selected terms are appropriate for the given domain and context (Lušicky and Wissik, 2015; Chiocchetti and Ralli, 2013; ISO 704:2009, 2009), use only synonyms and acceptable term variants (Lavagnino, 2010), and avoid common translation errors such as literal renderings and false friends (Lomaka, 2017). Rather than providing generic editing suggestions, Path 4 therefore equips the model with terminology-aware revision criteria grounded in established terminology management practice. Overall, this configuration tests whether structuring the post-editor's task with explicit, actionable criteria yields measurable gains over unconstrained post-editing.

Path 5: Rule PE + QA. Path 5 extends Path 4 with a Quality Assurance (QA) agent positioned downstream of the rule-guided post-editor. Quality assurance is, in fact, considered one of the main goals of terminology work (Schmitz and Straub, 2010). Terminology plays a role at every stage of the translation QA process, since it is one of the elements that must be checked against established standards (Popiolek, 2015). The QA agent verifies whether the post-editor applied all the rules and the glossary constraints. If all checks pass, the translation is accepted as the final translation for S5. If not, the post-edited translation is passed to the Supervisor agent, which provides textual feedback explaining to the post-editor what should be modified and how. The post-editor thus adjusts the translation based on the supervisor's feedback. This feedback paraphrase loop is repeated for a maximum of three iterations. At the end of the loop, a final attempt to repair the translation output is performed using a Post-Editor agent, originally powered by `gpt-4o-mini` and later replaced with `gpt-5.6-luna`.

This configuration tests the added value of an explicit, iterative self-correction mechanism over a single-pass rule-guided revision. The proposed multi-agent translation pipeline was inspired by the roles, activities, and workflow typically found in a professional translation setting. Rather than modelling agents as arbitrary computational components, each agent was designed to emulate a specific participant in the translation workflow. The Editor provides terminology resources or glossaries and the general translation guidelines and require-

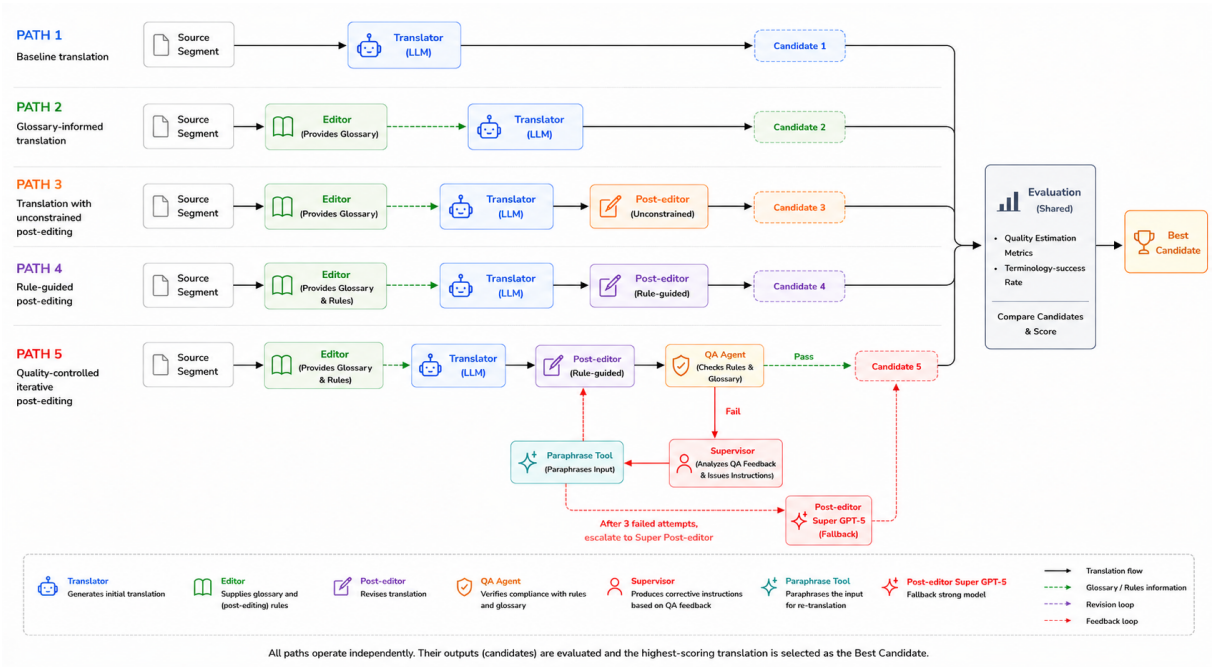

Figure 1: The UniOR-MT agentic pipeline. The pipeline is composed of five progressively enriched paths. For the mixed system submission, the best candidate is selected from the five paths in order to maximize quality estimation metrics.

ments; the Translator Agent corresponds to the professional translator responsible for producing the initial translation; the Post-Editor performs linguistic revision; the Quality Assurance Agent verifies compliance with predefined linguistic and terminological requirements; the Supervisor generates structured corrective feedback when quality issues are detected; and the Paraphraser Agent reformulates this feedback to facilitate subsequent refinement.

Consequently, the overall architecture reproduces the iterative nature of human translation workflows, where translation, revision, quality control, and feedback operate as complementary stages aimed at progressively improving translation quality. Although implemented through language models, the interaction among agents mirrors the collaborative practices commonly adopted in professional translation and localization environments.

3.2 System Implementation

The pipeline was implemented in Python using DSPy (Khattab et al., 2023), a framework for constructing modular programs powered by LLMs, through declarative task specifications. The translation, post-editing, and quality-assurance agents were implemented as separate DSPy modules, with `gpt-4o-mini` configured as the underlying language model through the OpenAI API, with temperature equal to 0.3. Prompts and module specifications were designed independently for each agent role and are reported in full in Appendix B.

3.3 Systems Submission

We submitted four systems: three individual configurations, namely i) ZS+T, ii) PE, and iii) Rule PE, and iv) one mixed system. The choice of individual submissions was based on development experiments conducted using the original pipeline implementation, in which all agents were powered by `gpt-4o-mini`. Under this implementation, the Rule PE+QA configuration did not provide sufficiently consistent improvements over the simpler paths and was therefore not selected as an individual submission.

Our primary system, called mixed, followed a broader candidate selection strategy. For each source segment, it selected the most suitable candidate that maximized a combined quality measure based on terminology accuracy and a document-level approximation of COMET-Kiwi (Rei et al., 2022), adapted to document-level evaluation, as described in Section 4.

Since the source documents exceed COMETKiwi's maximum context length, document-level quality estimation is approximated

by partitioning each document into contiguous windows containing approximately the same number of tokens. COMETKiwi is applied independently to each window, and the resulting scores are averaged to obtain a single quality estimate for the entire document. This last mixed system, therefore, combines the outputs of Path 1 to Path 5 in order to maximize the estimated quality.

After the standalone submission set had been determined, the final repair agent in the Rule PE+QA path was upgraded from `gpt-4o-mini` to `gpt-5.6-luna`. The stronger performance of Rule PE+QA in our internal evaluation therefore reflects the improved final configuration, whereas the individual submissions reflect the development results obtained with its original implementation.

4 Evaluation

Mode	Agentic path	TermAcc	docCOMET	SLIDE
English–Polish (en–pl)				
none	ZS	0.4620	0.0796	0.7488
	ZS+T	0.4619	0.0720	0.7562
	PE	0.4601	0.0806	0.7574
	Rule PE	0.4619	0.0780	0.7574
	Rule PE+QA	0.5001	0.1537	0.6608
proper	ZS	0.4620	0.0777	0.7488
	ZS+T	0.8836	0.0658	0.7464
	PE	0.8836	0.0724	0.7487
	Rule PE	0.8866	0.0698	0.7474
	Rule PE+QA	0.8895	0.1396	0.6741
random	ZS	0.4620	0.0803	0.7488
	ZS+T	0.4479	0.0601	0.7470
	PE	0.4479	0.0620	0.7482
	Rule PE	0.4479	0.0630	0.7481
	Rule PE+QA	0.4666	0.1026	0.7141
Spanish–Basque (es–eu)				
none	ZS	0.7048	0.0659	0.7297
	ZS+T	0.7126	0.0570	0.7333
	PE	0.7142	0.0587	0.7212
	Rule PE	0.7149	0.0567	0.7216
	Rule PE+QA	0.7494	0.0910	0.7228
proper	ZS	0.7008	0.0611	0.7306
	ZS+T	0.9068	0.0628	0.7297
	PE	0.9120	0.0647	0.7237
	Rule PE	0.9140	0.0640	0.7248
	Rule PE+QA	0.9235	0.0980	0.7288
random	ZS	0.6940	0.0666	0.7311
	ZS+T	0.7074	0.0454	0.7326
	PE	0.7148	0.0380	0.7212
	Rule PE	0.7136	0.0353	0.7258
	Rule PE+QA	0.7309	0.0530	0.7263

Table 1: Results for the English–Polish and Spanish–Basque language pairs.

We evaluate the proposed agentic paths internally using three metrics. TermAcc measures terminology adherence, while docCOMET (Vernikos et al., 2022) and SLIDE (Raunak et al., 2024) provide estimates of the overall translation quality at the document level. This analysis is conducted on a held-out test set derived from the training data and is intended to examine the contribution of the individual stages of the pipeline.

Table 1 reports results across the three terminology conditions and five agentic paths. The proper glossary produces the clearest gains in TermAcc for both language pairs. However, adding further processing stages after glossary injection does not result in proportional improvements. Differences among ZS+T, PE, and Rule PE are generally small, and the additional post-editing steps occasionally lead to lower scores for some translation quality metrics. This pattern suggests that the largest benefit is obtained from providing the correct terminology at translation time, while increasing the complexity of the subsequent revisions does not necessarily improve terminology adherence or overall translation quality across language pairs and evaluation metrics.

Rule PE+QA emerged as the strongest configuration according to docCOMET, consistently outperforming the other agentic paths across both language pairs and all terminology conditions. However, the same pattern is not consistently reflected by SLIDE. For English–Polish, Rule PE+QA obtains substantially higher docCOMET scores but lower SLIDE scores than the other configurations, whereas for Spanish–Basque the differences in SLIDE are smaller and less systematic. This divergence indicates that the two document-level metrics respond differently to the changes introduced by the QA stage, particularly for English–Polish.

Finally, the `random` condition shows that simply providing additional lexical information is not enough to improve performance. The gains observed with `proper` terminology appear to depend on the relevance and correctness of the supplied terms, while irrelevant terminology may even have a negative effect on translation quality.

The official evaluation on the blind test set is presented in Table 3 and 4 of Appendix C, and it supports the limited contribution of the additional processing stages. Overall, ZS+T achieves the strongest performance among our submitted systems, obtaining the best average scores for terminology success, COMET-22, XCOMET-XXL,

and MetricX-24, while Rule PE obtains a higher chrF++ score. The effect of post-editing appears to vary substantially across language pairs. For English-Polish, Rule PE and generic PE yield small improvements over ZS+T on some metrics. In contrast, for Spanish-Basque, both post-editing paths lead to lower terminology success and lower translation quality scores. Taken together, these results indicate that additional refinement stages may improve specific metrics or language pairs, but these gains are local and do not translate into a consistent advantage across evaluation conditions. Terminology injection therefore emerges as the most robust intervention in the proposed pipeline.

The greater degradation observed for Spanish–Basque after post-editing interventions may be related to the lower-resource nature of the Basque language. While additional post-editing can introduce useful revisions for a better-resourced target language such as Polish, the same generative interventions may be less reliable for Basque and may actually degrade the translations produced under direct terminology constraints.

5 Conclusions

In this paper, we presented UniOR-MT, an agentic pipeline for terminology-aware machine translation developed for the WMT 2026 Terminology Translation Shared Task. Our main contribution is an incremental evaluation framework that isolates the effect of successive terminology interventions within a unified translation workflow. This design allows us to quantify how each stage contributes to terminology success rate and overall translation quality, providing insights into the trade-offs between increasingly sophisticated terminology-aware strategies.

Our experiments show that providing a bilingual glossary substantially improves terminology adherence, whereas generic and rule-guided post-editing yield comparatively smaller gains. This indicates that increasing the number of processing stages after glossary injection does not necessarily translate into proportional improvements in either terminology adherence or document-level translation quality. In our internal evaluation, the largest improvements in docCOMET are achieved by introducing an explicit quality assurance stage with iterative feedback. However, these gains are not consistently reflected by SLIDE, highlighting that additional refinement may improve some dimensions of translation quality while negatively affecting others. The official blind evaluation further shows that ZS+T achieves the strongest overall performance among our submitted systems. While post-editing produces small improvements for English–Polish on some metrics, it substantially degrades performance for Spanish–Basque. Notably, these results were obtained with a pipeline relying primarily on `gpt-4o-mini`, suggesting that structured terminology-aware workflows can achieve strong performance without relying on larger language models throughout the pipeline.

Overall, our findings suggest that the effectiveness of terminology-aware MT depends not only on the availability of lexical resources but also on how these resources are integrated into a structured revision workflow. Increasing pipeline complexity does not consistently lead to better performance: while accurate terminology resources provide clear benefits, subsequent refinement stages yield more limited, metric-dependent, and language-pair-dependent gains. By progressively introducing terminology-aware components within a unified pipeline, our evaluation framework provides a more interpretable analysis of their individual contribution to terminology adherence and overall translation quality than evaluating complete end-to-end pipelines alone.

6 Limitations

The proposed pipeline was evaluated exclusively on the WMT 2026 Terminology Translation Shared Task (Track 1), covering two language pairs and a limited number of specialised domains. Consequently, the observed improvements may not generalise to other languages, domains, or translation settings. Furthermore, the pipeline was implemented using only a small selection of language model, meaning that the reported gains primarily reflect the contribution of the agentic workflow and prompting strategy rather than model diversity. Finally, our internal quality estimation relied on an approximation of COMETKiwi obtained by averaging scores over fixed-length context windows. Although this enabled document-level comparison across systems, it may not fully capture discourse phenomena spanning longer contexts. Future work should validate these findings through human evaluation and investigate stronger specialised agents and long-context quality estimation methods.

Acknowledgments

This work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples "L'Orientale", through a doctoral grant (ID 39-411-24-DOT23A27WJ-6603) to the first author, established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan. This work was also partially supported by the PhD programme in Humanities and Technologies funded by the University of Macerata under D.R. No 253/2023. This work has been partially supported by the 4MLP Project, funded by the University of Naples "L'Orientale" under the University Research Projects funding scheme (PRA 2025) - CUP C66125002860005.

References

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Lynne Bowker. 2015. Terminology and translation. In Hendrik J. Kockaert and Frieda Steurs, editors, *Handbook of Terminology: Volume 1*, pages 304–323. John Benjamins, Amsterdam/Philadelphia.

Vicent Briva-Iglesias. 2025. Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 365–377, Geneva, Switzerland. European Association for Machine Translation.

Vicent Briva-Iglesias and María Ferre-Fernández. 2026. AI-assisted cultural heritage dissemination: Comparing NMT and glossary-augmented LLM translation in rock art documents. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, Tilburg, Netherlands. European Association for Machine Translation.

Maria Teresa Cabré. 2010. Terminology and translation. In Yves Gambier and Luc van Doorslaer, editors, *Handbook of Translation Studies: Volume 1*, pages 356–365. John Benjamins, Amsterdam/Philadelphia.

Adrian Charkiewicz, Pinzhen Chen, Thierry Etchegoyhen, Harritxu Gete Ugarte, Kamil Guttmann, Xu Huang, David Ponce, Artur Nowakowski, Frédéric Odermatt, Arturo Oncevay, Kirill Semenov, Dawei Zhu, and Vilém Zouhar. 2026. Findings of the WMT26 terminology translation task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.

Elena Chiocchetti and Natascia Ralli. 2013. *Guidelines for Collaborative Legal/Administrative Terminology Work*. Eurac Research, Bolzano.

Keiran J. Dunne. 2013. Translation tools. In Carol A. Chapelle, editor, *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell Publishing, Malden, MA.

María Fernández-Parra. 2025. *Terminology Management for Translators*. Routledge, London.

Ivan Grubišić and Damir Korencic. 2025. IRB-MT at WMT25 terminology translation task: Metric-guided multi-agent approach. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1302–1334, Suzhou, China. Association for Computational Linguistics.

Kamil Guttmann, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 terminology translation task: A multi-objective reranking strategy for terminology-aware translation via Pareto-optimal decoding. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1276–1283, Suzhou, China. Association for Computational Linguistics.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 2*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Iikka Hauhio and Théo Friberg. 2024. Mitra: Improving terminologically constrained translation quality with backtranslations and flag diacritics. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation: Volume 1*, pages 100–115, Sheffield, UK. European Association for Machine Translation.

Pengcheng Huang, Yongyu Mu, Yuzhang Wu, Bei Li, Chunyang Xiao, Tong Xiao, and Jingbo Zhu. 2024. Translate-and-revise: Boosting large language models for constrained translation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics: Volume 1*, pages 1059–1074, Taiyuan, China. Chinese Information Processing Society of China.

ISO 11669:2024. 2024. Translation projects – guidance. Technical report, International Organization for Standardization, Geneva.

ISO 17100:2015. 2015. Translation services – requirements for translation services. Technical report, International Organization for Standardization, Geneva.

ISO 704:2009. 2009. Terminology work – principles and methods. Technical report, International Organization for Standardization, Geneva.

Akshat Jaswal. 2025. It takes two: A dual stage approach for terminology-aware translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1344–1350, Suzhou, China. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Elisa Lavagnino. 2010. L'instabilità dei termini complessi all'interno delle LSP: un confronto italiano/francese. *Publifarum*, (12).

Anzhelika O. Lomaka. 2017. Literal approach to translation: A classification and literature review. *XLinguae*, 10(4):234–244.

Vesna Lušicky and Tanja Wissik. 2015. *Procedural Manual on Terminology: Translation-oriented Terminology Work*. Secretariat for European Affairs.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Andrew Y. Ng. 2025. Translation-agent: A multi-agent translation framework. `https://github.com/andrewyng/translation-agent`. GitHub repository.

Silvia Pavel and Diane Nolet. 2001. *Handbook of Terminology*. Public Works and Government Services Canada, Translation Bureau, Ottawa.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. Adding chocolate to Mint: Mitigating metric interference in machine translation. *Transactions of the Association for Computational Linguistics*, 13:1319–1339.

Monika Popiolek. 2015. *Terminology Management within a Translation Quality Assurance Process*. John Benjamins, Amsterdam/Philadelphia.

Fernando Prieto Ramos. 2021. Ensuring consistency and accuracy of legal terms in institutional translation: The role of terminological resources in international organizations. In *Institutional Translation and Interpreting: Assessing Practices and Managing for Quality*. Taylor & Francis, London.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 205–211, Mexico City, Mexico. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hanna Risku. 2006. Cooperation and quality assurance in technical translation projects. *Language at Work – Bridging Theory and Practice*, 1(1).

Federica Scarpa. 2001. *La Traduzione Specializzata: Lingue Speciali e Mediazione Linguistica*. Hoepli, Milano.

Klaus-Dirk Schmitz and Daniela Straub. 2010. *Erfolgreiches Terminologiemanagement im Unternehmen: Praxishilfe und Leitfaden: Grundlagen, Umsetzung, Kosten-Nutzen-Analyse, Systemübersicht*. TEC More GmbH.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

King-kui Sin, Xi Xuan, Chunyu Kit, Clara Ho-yan Chan, and Honic Ho-kin Ip. 2025. Solving the unsolvable: Translating case law in Hong Kong. *arXiv preprint arXiv:2501.09444*.

Maria Carmen Staiano, Johanna Monti, Francesca Chiusaroli, Slava Tykhonov, Ken Hawkins, Hélène Jacot des Combes, Saini Yang, Urbano Fra Paleo, and Virginia Murray. 2026. Benchmarking multilingual terminology translation for UNDRR-ISC hazard information profiles. In *Proceedings of the Conference on New Trends in Translation and Interpreting Technology*, pages 207–230, Dubrovnik, Croatia. INCOMA Ltd.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level mt metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, and Yang Liu. 2021. Neural machine translation with explicit phrase alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1001–1010.

A Explicit Post-Editing Rules

1. **Glossary conformity.** Every glossary pair must be respected. Wherever a source term appears in the source sentence, the translation must use the corresponding target term exactly, allowing only the grammatical inflection required by the target language. If the draft uses a synonym or paraphrase, replace it with the prescribed glossary term.

2. **Terminological consistency.** Use the same target term for the same concept throughout the sentence. Do not alternate between the glossary term and a synonym.

3. **Domain adequacy.** Use renderings appropriate to the relevant domain, such as enterprise software or technical documentation. Prefer established domain-specific terminology to general-language wording.

4. **Correct grammatical variants.** Inflect glossary terms correctly according to the grammar of the target language. Case, number, gender agreement, orthographic variants, abbreviations, and acronyms must all be well formed. Adapting the grammatical form of a glossary term to fit the sentence is required and does not constitute a violation of Rule 1.

5. **Do-not-translate terms.** Terms that must remain in the source language, including product names, brand names, user-interface commands, code identifiers, and file paths, must be preserved exactly as they appear in the source. If the glossary maps a term to itself, treat it as a do-not-translate term.

6. **Avoidance of word-for-word literalism.** Except for constrained glossary terms, translate the meaning rather than the surface form. Restructure the sentence whenever required by the target language so that the translation reads naturally and fluently.

7. **Avoidance of false friends.** Check words that are formally similar across the source and target languages but differ in meaning. Ensure that each term conveys the intended source meaning rather than that of a misleading look-alike.

8. **Minimal editing.** Make only the changes required by the preceding rules or those necessary to correct clear problems of accuracy, grammar, or fluency. Do not rewrite acceptable phrasing, and do not add or omit information relative to the source.

B Prompts Configuration for the Agents

Path 1: Baseline Translator

Translator agent

Translate the source sentence while enforcing every terminology pair exactly, allowing only grammatical inflection. If no terminology is provided, produce an accurate and fluent translation.

Path 2: Translator with Glossary

Translator agent

Translate the source sentence into the target language. You MUST use the provided terminology: for every source-term-to-target-term pair, the translation must render the source term using exactly the specified target term, inflected only as required by target-language grammar. If the terminology is "(none)", translate the sentence accurately and fluently. Output only the translation.

Path 3: Translator and Generic Post-Editor

Post-editor agent

Post-edit the draft translation. Ensure that every terminology constraint is respected exactly, allowing only grammatical inflection, and correct any accuracy, grammar, or fluency issues. If no terminology is provided, focus exclusively on translation quality. Keep edits minimal and output only the final translation.

Path 4: Translator and Rule-Based Post-Editor

Rule-based post-editor agent

Revise the draft translation by applying the complete rule set in addition to enforcing the glossary. The rules require glossary conformity, terminological consistency, domain adequacy, correct grammatical inflection, preservation of protected terms, natural target-language phrasing, avoidance of false friends, and minimal editing. Output only the final translation.

Path 5: Full Agentic Pipeline

Quality-assurance agent

Verify that the translation complies with every rule in the rule set and with every terminology constraint. Check the rules individually against the translation. Set `passed=True` only if all requirements are satisfied. If any requirement is violated, set `passed=False`; in the feedback, identify each violated rule and state concretely how the translation should be revised to enforce it.

Supervisor agent

Diagnose why the translation fails the QA check and produce concrete, actionable feedback for a post-editor to fix it. Ground every point in a specific rule or terminology constraint and state exactly what to change.

Paraphrase agent

Rephrase and sharpen the guidance instead of repeating it verbatim, calling out whatever the earlier wording likely left ambiguous or the post-editor missed.

Table 2: Prompt configurations used in the proposed translation pipeline

C Official Shared Task Evaluation

Table 3 reports the official WMT 2026 evaluation scores for our submitted systems across the three terminology conditions.

System	Proper			Random			Noterm		
	Term	**chrF++**	**COMET**	**Term**	**chrF++**	**COMET**	**Term**	**chrF++**	**COMET**
UniOR-ZSL+Term	**86.4**	68.8	**83.9**	**45.5**	65.4	**81.5**	**48.7**	66.0	**83.0**
UniOR-RulePE	84.7	68.9	82.6	45.0	65.8	80.5	47.4	66.4	81.2
UniOR-GenPE	84.1	68.8	82.5	44.5	66.0	80.2	47.1	66.4	81.4
UniOR-Mixed	76.4	**70.4**	76.7	41.6	**67.7**	75.0	47.1	**67.5**	78.3

Table 3: Official evaluation results for the UniOR submissions in Track 1 across the `proper`, `random`, and `noterm` terminology conditions. Term denotes lemmatized terminology success (%), while chrF++ and COMET refer to document-level chrF++ and COMET-22, respectively.

System	chrF++			Term Success (lemma)			COMET-22			XCOMET-XXL			MetricX-24 XL ↓		
	Avg.	en–pl	es–eu	Avg.	en–pl	es–eu	Avg.	en–pl	es–eu	Avg.	en–pl	es–eu	Avg.	en–pl	es–eu
ZS+T	68.8	73.4	64.1	**86.4**	87.9	**84.9**	**83.9**	87.9	**79.9**	**75.8**	84.0	**67.6**	**5.60**	4.72	**6.48**
PE	68.8	73.4	64.2	84.1	87.7	80.5	82.5	**88.0**	77.0	73.9	**84.2**	63.7	6.00	**4.66**	7.33
Rule PE	68.9	73.5	64.2	84.7	**88.4**	81.0	82.6	87.9	77.3	74.2	84.1	64.3	5.95	4.69	7.21
Mixed	**70.4**	**74.5**	**66.3**	76.4	70.8	81.9	76.7	74.6	78.8	65.8	64.5	67.0	7.79	8.97	6.61

Table 4: Official evaluation results for the UniOR subsmissions. The best result in each column is shown in bold. Lower values are better for MetricX-24 XL.