ICB-UMA at #SMM4H–HeaRD 2026: Hybrid Clinical Entity Projection for MultiClinAI: Adaptive Candidate Windows, XGBoost, and LLM Refinement

Álvaro Rey-Blanes[1,2], Sara Giménez-Gómez[1,2], Francisco J. Veredas[1,2], Francisco J. Moreno-Barea[1,2]

[1]Department of Programming Languages and Computer Sciences, Universidad de Málaga Bulevar Louis Pasteur, 35, 29071 Málaga, Spain

[2]Research Institute of Multilingual Language Technologies, Universidad de Málaga C/ Severo Ochoa, 4, Málaga TechPark, Campanillas, 29071 Málaga, Spain

Abstract

This paper presents our submission to the MultiClinAI Shared Task (Gallego-Donoso et al., 2026) on cross-lingual clinical entity annotation projection from Spanish to English. Our system transfers expert annotations for *Diseases*, *Symptoms* and *Procedures* entities. The approach integrates three core components: adaptive candidate window generation, an XGBoost classifier leveraging surface and semantic features, and an LLM-based post-processing stage to resolve complex misalignments. Our highest-performing run ranked 3rd on the official leaderboard, achieving strict F1 scores of 0.737, 0.549, and 0.538 for *Diseases*, *Procedures*, and *Symptoms*, respectively. These results show that combining supervised candidate scoring with targeted LLM refinement provides a robust strategy for clinical entity projection.

1 Introduction

Named Entity Recognition (NER) is a foundational pillar in clinical Natural Language Processing (NLP), allowing the automatic structured identification of critical clinical variables such as diagnoses, symptoms, and medical procedures from unstructured Electronic Health Records (EHRs) (Névéol et al., 2018). However, building high-quality NER systems requires a large corpus annotated by clinical experts. Despite the global nature of medical knowledge, the development of these robust NER models for non-English languages, such as Spanish, often faces significant resource constraints (Miranda-Escalada et al., 2020; Moreno-Barea et al., 2025). Manual annotation by experts is not only economically prohibitive but also time-intensive, creating a "data bottleneck" that slows the deployment of cross-border medical research.

Unlike traditional NER, where a model is trained directly on target-language data, annotation projection leverages the semantic equivalence between parallel texts, offering a cost-effective alternative (Yarowsky et al., 2001; Ehrmann et al., 2011). This process involves taking the manually validated entity mentions produced by domain experts in a high-resource language and precisely mapping their character offsets onto comparable texts in a lower-resource language. This approach seeks to avoid the need for manual annotation from scratch in the target language by using: Machine Translation, to preserve the contextual integrity of the clinical narrative (Ni et al., 2017; Jain et al., 2019); and Entity Alignment Strategies, to ensure that complex medical terms (e.g., *infarto agudo de miocardio*) are mapped accurately to their English equivalents (*acute myocardial infarction*) despite differences in syntax and word order (Yousef et al., 2023).

The MultiClinAI shared task addresses this disparity by exploring the automatic generation of comparable multilingual corpus through named entity projection. MultiClinAI provides a multilingual benchmark with expert annotations in a high-resource source language (Spanish) to target different target languages (English, Italian, Dutch, Swedish, Romanian and Czech), inviting teams to perform cross-lingual entity projection (Gallego-Donoso et al., 2026; Lopez-Garcia et al., 2026).

In this paper, we evaluate the efficacy of various alignment, Machine Learning (ML) and Large Language Model (LLM) based strategies for the Spanish-to-English projection of clinical entities. The system submitted in the MultiClinAI shared task combines three components: an adaptive windowing strategy that generates and filters candidate English spans for each annotated Spanish entity; a supervised XGBoost binary classifier (Chen and Guestrin, 2016) that scores each ⟨Spanish entity, English candidate span⟩ pair using surface and semantic similarity features; and an LLM-based correction stage that handles cases where no satisfactory match is found, querying a LLM (Gemma Team, 2024) against the full English document to locate the equivalent expression in the target text.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 127–132

Table 1: Corpus statistics by entity type and split. The test partition corresponds to the full set released by the organisers; the subset used for official scoring is an internal holdout not disclosed to participants.

	Documents		Entities		Entities per Doc.		Windows		Wind. per Doc.		Wind. per Entity	
Type	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Disease	881	3 260	12 903	59 552	14.65	18.27	106 206	409 921	120.55	125.74	8.23	6.88
Procedure	825	3 260	14 055	63 481	17.04	19.47	113 864	447 515	138.02	137.27	8.10	7.05
Symptom	850	3 259	13 545	81 510	15.94	25.01	112 394	487 591	132.23	149.61	8.30	5.98

2 Data and Methods

2.1 Corpus

The shared task provides a set of clinical narratives in Spanish and English validated by domain experts, forming the gold standard for the projection task (Gallego-Donoso et al., 2026). Each Spanish document is paired with its corresponding English counterpart, and the annotations cover three semantic categories: *Diseases*, *Symptoms*, and *Procedures*. A preliminary analysis of the corpus revealed that approximately 30% of the document pairs exhibited non-trivial structural misalignment between source and target texts, complicating direct offset projection and motivating our window-based candidate generation strategy.

Table 1 summarises the corpus statistics by entity type and split. The training set comprises 825-881 documents—depending on entity type— with between 12,903 and 14,055 annotated entities. The test set is considerably larger, containing 3,259-3,260 documents and 59,552 to 81,510 entities, reflecting a ratio of approximately 1:4 across train and test. It is important to note that the test set described here corresponds to the full set of documents released by the shared task organisers; the subset used to compute the official leaderboard results is an internal holdout whose composition is not disclosed to participants. Consequently, the results reported in Section 3 are evaluated against this unknown subset rather than the complete test partition shown in Table 1. Window generation yields approximately 8 candidate spans per entity across all types and splits (ranging from 5.98 to 8.30), while it is produced a lower *windows per entity* ratio in the test set, consistent with the higher entity density observed in that partition.

2.2 Candidate Window Generation

Given an annotated entity span in a Spanish document, the goal of candidate generation is to identify a set of English character spans likely to contain the corresponding entity. We adopted an adaptive window strategy over the English text. Let s be the token length of the source entity span. We defined a primary search window of $[s - 4,\ s + 4]$ tokens, allowing for morphological and syntactic variation between languages.

To account for systematic length differences between Spanish and English, we applied a Language Expansion Factor (LEF) of 0.95, modelling the empirical observation that English translations of Spanish clinical text tend to be approximately 5% shorter. The candidate search is anchored around a predicted target offset computed by scaling the source offset by the LEF, with a tolerance margin of ± 100 tokens to absorb residual misalignment.

Window boundaries are constrained to avoid splitting at linguistically incoherent positions: expansion halts if it reaches a stopword or a punctuation token that would produce a truncated entity mention. This heuristic reduces the number of noisy candidates without discarding valid projections.

2.3 Feature Extraction

For each $\langle$*source entity, target candidate*$\rangle$ pair we extracted two families of features.

Surface features. Character-level and token-level similarity metrics capturing shallow overlap between the Spanish entity and the English candidate: Levenshtein edit distance (Levenshtein, 1966), Jaccard coefficient over character n-grams ($n \in \{3, 4\}$), TF-IDF cosine similarity, character and token length ratios, and the offset difference between the expected and observed positions in the target document.

Semantic features. We encoded the source entity and the target candidate independently using the `paraphrase-multilingual-mpnet-base-v2` SBERT model (Reimers and Gurevych, 2019), a 278M-parameter transformer that maps text into a 768-dimensional space optimised for cross-lingual similarity. The cosine similarity between the two resulting vectors serves as a semantic alignment feature.

Table 2: Results obtained from the validation subset of the training split by entity type, including corpus statistics.

Entity	Windows	Entities	Docs	Wind. per Entity	P	R	F1
Disease	21 222	2 581	761	8.22	0.8191	0.8160	0.8175
Procedure	22 772	2 811	750	8.10	0.8517	0.7947	0.8222
Symptom	22 550	2 709	737	8.32	0.8325	0.7505	0.7894

2.4 Pair Scoring with XGBoost

We framed candidate selection as a binary classification problem: given a ⟨source, candidate⟩ pair and its feature vector, the model predicts whether the candidate is the correct projection of the source entity. We trained an XGBoost classifier (Chen and Guestrin, 2016) on positive pairs (annotated matches) and negative pairs (all other candidates for the same source entity).

The training set exhibits substantial class imbalance, as each source entity typically generates approximately 8 candidate windows but only one is correct, resulting in a positive rate of 12%. To address this, we set the `scale_pos_weight` hyperparameter to $\sqrt{N^-/N^+}$, where N^- and N^+ are the counts of negative and positive training examples respectively. This conservative weighting improves recall on the minority class without overly penalising precision. The XGBoost model was configured with the hyperparameters listed in Table A1.

2.5 LLM-based Post-processing

Candidate pairs predicted as positive by the XGBoost classifier were subsequently evaluated by a first-pass LLM semantic evaluator, which assigned one of three labels: *exact match* (the English window is a direct or clearly equivalent translation of the Spanish entity), *mid match* (the correspondence is partial, approximate, or contextually related), or *no match* (no semantic correspondence). To ensure conservative acceptance, the evaluator was explicitly instructed to prefer *mid match* over *exact match* in ambiguous cases, routing borderline pairs to the correction stage rather than accepting them silently.

Pairs labelled *mid match* or *no match* were routed through a correction stage. For each such pair, the pipeline loaded the complete English document and queried a LLM (Gemma 4 31B, served locally via Ollama[1]) to suggest a more faithful English realisation of the Spanish entity mention. The suggestion is then located within the English document using exact string search; in cases of multiple occurrences, the occurrence nearest to the expected target offset is selected. Validated suggestions are reintegrated into the annotation output. This step extends coverage for difficult entity types and structurally misaligned document pairs. (See Appendix B for full prompts.)

[1] https://ollama.com

3 Results

3.1 Validation Split Results

The available data was split into 80% for training and 20% for validation in order to perform hyperparameter tuning of the ML models. The split was carried out at the entity level; that is, all candidate windows belonging to the same source entity are assigned exclusively to either the training or the validation partition, ensuring that no entity seen during training appears in the validation set, thereby preventing any information leakage.

Table 2 reports the resulting statistics and XGBoost performance on the validation partition. The classifier achieves F1 scores above 0.78 across all entity types, with *Procedure* yielding the highest score (F1 = 0.822) and *Symptom* the lowest (F1 = 0.789), consistent with the greater morphological variability of symptom mentions.

3.2 Official Task Results

We submitted four runs to the shared task evaluation. Run 3 and Run 4 consistently outperformed Run 1 and Run 2 across all entity types and metrics; we therefore focus the analysis on these two configurations. The key difference between Run 3 and Run 4 is the number of entity pairs processed by the XGBoost scorer (Chen and Guestrin, 2016): Run 4 includes approximately 12,000 additional candidate pairs with respect to Run 3, reflecting a wider window-generation sweep.

Table 3 reports strict and character-level results for the *Disease* entity type, where per-run breakdowns are available. Table 4 summarises performance for *Procedure* and *Symptom*. The system ranked **3rd** in the shared task leaderboard across all three entity types under strict F1. *Disease* entities yielded the strongest results (strict F1 = 0.737, char F1 = 0.852), benefiting from their relatively higher surface-form overlap between Spanish and English.

Table 3: Results on *Disease* per submitted run (strict span match and character-level overlap). Rank refers to the shared task leaderboard position by strict F1.

	Strict			Char-level			
Run	**P**	**R**	**F1**	**P**	**R**	**F1**	**Rank**
Run 4	0.738	0.735	**0.737**	0.865	0.839	**0.852**	3rd
Run 3	0.738	0.733	0.736	0.865	0.837	0.851	4th
Run 2	0.689	0.689	0.689	0.830	0.806	0.817	5th
Run 1	0.415	0.411	0.413	0.770	0.750	0.760	6th

Table 4: Best-run results per entity type (strict span match and character-level overlap). Rank refers to the shared task leaderboard position by strict F1.

	Strict			Char-level			
Entity	**P**	**R**	**F1**	**P**	**R**	**F1**	**Rank**
Disease	0.738	0.735	**0.737**	0.865	0.839	**0.852**	3rd
Procedure	0.552	0.545	**0.549**	0.724	0.681	**0.702**	3rd
Symptom	0.540	0.537	**0.538**	0.743	0.701	**0.721**	3rd

Procedure and *Symptom* entities proved more challenging, with strict F1 scores of 0.549 and 0.538 respectively, a gap attributable to greater morphological variation and the shorter, more ambiguous surface forms typical of symptom mentions.

Comparing Run 3 and Run 4 on *Disease*, the broader candidate sweep of Run 4 (+12,000 entity pairs) yielded a marginal but consistent gain in recall (0.733 $\rightarrow$ 0.735) without degrading precision (0.738 in both runs), confirming that a wider search horizon improves coverage at negligible cost to exactness. The character-level metrics are notably higher than strict metrics across all entity types (e.g. char F1 = 0.852 vs. strict F1 = 0.737 for *Disease*), indicating that the system correctly identifies the relevant text region in most cases even when the exact span boundaries do not match the gold annotation accurately.

3.3 LLM-based Post-processing Impact

The LLM-based post-processing pipeline operates as follows: XGBoost-positive predictions are assessed by a semantic evaluator. Pairs classified as *mid match* or *no match* are then routed through a correction stage. Across the test set, 14,842 entity pairs were routed to the correction stage, representing targeted LLM intervention for difficult cases where surface-level and semantic features alone were insufficient.

3.4 Systematic Failure Analysis

A detailed analysis of the validation predictions reveals several systematic failure modes. First, morphological variability, where Spanish entity mentions are significantly shorter or longer than their English equivalents, causes window-based candidates to rank poorly despite semantic correspondence, resulting in high length differences (DIFF_NUM_WORDS up to ±11) that penalise surface-level similarity features. Second, clinical abbreviations and acronyms present a distinct challenge. Acronyms refer to identical clinical concepts but exhibit low lexical overlap (Levenshtein $\approx$ 0.20), causing the XGBoost classifier to assign low scores unless intervened by the LLM. Third, word order reordering between languages creates misalignments. Spanish phrases like *queratitis ulcerativa periférica superior* may correspond to English *superior peripheral ulcerative keratitis*, where word order differs significantly. This variation is partially captured by semantic features (SBERT similarity), but surface-level metrics alone prove insufficient. These findings suggest improvements should focus on language-aware length normalisation, clinical abbreviation expansion or matching, and word-order-invariant features.

4 Discussion and Future Work

The experimental results confirm that a combination of adaptive windowing, heterogeneous features, and a discriminative pair scorer constitutes an effective strategy for cross-lingual annotation projection in the clinical domain. The LEF and the token-margin mechanism proved particularly valuable for handling the $\sim$30% of document pairs affected by structural misalignment, cases where direct offset transfer would fail entirely.

Several directions remain open for future work. First, the binary classification framing could be replaced by a Learning-to-Rank formulation, using a ranker such as XGRanker [2], to directly optimise the ranking of candidate spans rather than a fixed decision threshold, which is expected to benefit entity types where multiple candidates receive similar scores. Second, the pipeline was evaluated exclusively on the Spanish–English pair; extending it to the other MultiClinAI language pairs (Czech, Romanian, Italian and Swiss German) requires minimal adaptation of the LEF and stopword lists. Third, a principled feature selection procedure such as SHAP-based importance (Lundberg and Lee, 2017) could further improve generalisation by removing redundant or noisy features from the XGBoost input.

[2] https://xgboost.readthedocs.io/en/stable/tutorials/learning_to_rank.html

5 Limitations

The system was developed and evaluated exclusively on the Spanish-English language pair, and its performance on the remaining MultiClinAI language pairs (Czech, Romanian, Italian, and Swiss German) remains untested. The LLM-based post-processing stage relies on Gemma 4 31B (Gemma Team, 2024) served locally via Ollama, which requires substantial computational resources and may limit reproducibility in low-resource settings. Finally, the current evaluation does not include a fine-grained error analysis across document types or entity subtypes, which would help identify systematic failure modes of the pipeline.

Data and Code Availability

Data is publicly available at MultiClinAI Shared Task Training Data (Zenodo). Code can be found at ICB-MultiClinCorpus Repository (GitHub)

Acknowledgments

The authors would like to thank the organizers of the MultiClinAI shared task and the Social Media Mining for Health and Health Real-World Data for providing the corpus and evaluation framework. The authors acknowledge the support from the Ministerio de Ciencia e Innovación (MICINN) under project PID2024-155334OB-I00.

References

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 118–124.

Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Google DeepMind Gemma Team. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estàpe, and Martin Krallinger. 2020. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at CodiEsp track of CLEF eHealth 2020. In *CLEF 2020 Working Notes*.

Francisco J Moreno-Barea, Guillermo López-García, Héctor Mesa, Nuria Ribelles, Emilio Alba, José M Jerez, and Francisco J Veredas. 2025. Named entity recognition for de-identifying spanish electronic health records. *Computers in Biology and Medicine*, 185:109576.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: Opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023. Named entity annotation projection applied to classical languages. In *Proceedings of the 7th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, pages 175–182.

A XGBoost Training Hyperparameters

Table A1: XGBoost classifier hyperparameters used for entity pair scoring.

Hyperparameter	Value
n_estimators	500
max_depth	4
learning_rate	0.05
subsample	0.9
colsample_bytree	0.9
reg_lambda	2.0
reg_alpha	0.1
min_child_weight	5
gamma	0.0
max_delta_step	1
scale_pos_weight	$\sqrt{N^-/N^+}$
objective	binary:logistic
eval_metric	aucpr
random_state	42

B Prompts

Evaluation Prompt

You are a strict bilingual clinical text matcher.
Your task is to determine whether a Spanish entity mention is represented in an English text window. If there is any noise, consider it a no match. Spanish entity:
`{src_entity_text}`
English window:
`{tgt_window_text}`
Return a JSON object with this exact schema: {{"match_label": "exact match | mid match | no match", "justification": "string"}}
Label definitions: - exact match: the English window clearly contains the same concept or a direct translation of the Spanish entity - mid match: partial, approximate, broader, narrower, abbreviated, or contextually related match - no match: the English window does not represent the same concept
Rules: - Compare meaning, not literal wording - Consider translations, abbreviations, paraphrases, and morphological variants - If uncertain between exact match and mid match, choose mid match - If match_label is "exact match", justification must be an empty string - If match_label is "mid match" or "no match", justification must contain a brief explanation - Return JSON only

Search Prompt

You are a biomedical NER expert. Below is a clinical document in English, followed by a list of entity spans that were incorrectly extracted. For each entity, using the document context and the Spanish reference, suggest the most accurate English entity string from the document.
DOCUMENT
`{document_text}`
INCORRECT ENTITIES:
`{entity_lines}`
Respond ONLY with a JSON array, no explanation:
{"id": 1, "suggestion": "..."}