SINAI at GRACE 2026 Track 1: GlobalPointer with GSSDAF Span Interaction and Majority Voting Ensemble for Argument Mining in Clinical Trial Abstracts

Lucas Molino Piñar[1,*,†], Manuel Carlos Díaz Galiano[1,†] and María Teresa Martín Valdivia[1,†]

[1]SINAI Research Group, Computer Science Department, Universidad de Jaén, Spain

Abstract

This paper describes the system submitted by the SINAI team to Track 1 of the GRACE shared task at IberLEF 2026. The task addresses Argument Mining over Spanish abstracts of Randomised Controlled Trials (RCTs), encompassing two hierarchical subtasks: argumentative component detection (Subtask 1) and directed relation classification between components (Subtask 2). Our approach builds on the GlobalPointer span-extraction framework, enhanced with a novel *Global Span Semantic Dependency-Aware Filtering* (GSSDAF) interaction module that refines the span logit matrix through local convolutional and global axial-attention branches. Rotary Position Embeddings (RoPE) provide position-aware representations in the biaffine scoring space. For relation classification, we employ a span-pair classifier that combines mean-pooled span representations with entity-type and logarithmic-distance embeddings, trained with α-balanced Focal Loss. Both heads share a single BERT-based encoder (BETO). A 5-fold cross-validated ensemble with character-level majority voting consolidates predictions at inference time. Our system achieved an Overall score of **26.42** on the official Track 1 leaderboard (Subtask 1 Strict F1: 36.87, Subtask 2 Strict F1: 15.97).

Keywords

Argument Mining, Clinical Trials, GlobalPointer, GSSDAF, Span Extraction, Relation Classification, Ensemble Learning, IberLEF 2026

1. Introduction

Argument Mining (AM) is the automatic identification and classification of argumentative structures within text [1]. When applied to the biomedical domain, AM can assist clinicians and systematic reviewers in navigating the vast landscape of clinical evidence by surfacing the key claims, premises, and evidential relations embedded in research publications [2].

Randomised Controlled Trials (RCTs) represent the gold standard of clinical research methodology. Their structured nature, comparing an intervention group against a control group under randomised allocation, produces abstracts with particularly transparent argumentative patterns, making them an ideal testbed for AM systems [2].

The GRACE shared task at IberLEF 2026 [3], organised within the IberLEF 2026 evaluation forum [4], addresses AM over Spanish RCT abstracts based on the Spanish AbstRCT Dataset [5] and ClinArgES dataset [6]. Track 1 is organised into two hierarchical subtasks:

(i) **Subtask 1 — Component Detection**: Identify and classify the boundaries of argumentative components (Premise, Claim, MajorClaim) at the character level within the abstract.

(ii) **Subtask 2 — Relation Detection**: Classify directed relations (Support, Attack, Partial-Attack) between pairs of previously detected components.

Traditional sequence-labelling approaches (e.g., BIO tagging with CRFs) struggle to capture the nested and overlapping nature of argumentative spans, and they require a separate mechanism for

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

lmolino@ujaen.es (L. M. Piñar); mcdiaz@ujaen.es (M. C. D. Galiano); maite@ujaen.es (M. T. M. Valdivia)

0009-0009-5990-9802 (L. M. Piñar); 0000-0001-9298-1376 (M. C. D. Galiano); 0000-0002-2874-0401 (M. T. M. Valdivia)

boundary detection [7]; in contrast, span-extraction methods that score all candidate $(start, end)$ pairs in a biaffine matrix offer a more natural formulation for this problem. In this paper, we present the SINAI system[1], which addresses this challenge through several key contributions: a joint architecture with a shared BERT encoder and dual task-specific heads that enables parameter-efficient multi-task learning; a GlobalPointer + GSSDAF entity head that models inter-span dependencies through local (convolutional) and global (axial-attention) branches with adaptive gating applied directly on the span logit matrix; a span-pair relation classifier that leverages mean-pooled span representations enriched with entity-type and logarithmic-distance embeddings, trained using α-balanced Focal Loss to address severe class imbalance; an Optuna-based hyperparameter optimisation pipeline; and a 5-fold ensemble with character-level majority voting that consolidates entity and relation predictions. Ultimately, our system achieved highly competitive results, ranking **1st** on the official Track 1 leaderboard.

The remainder of this paper is organised as follows: first, we describe the task in the following section. Section 3 details our system architecture. Section 4 describes the experimental setup. Section 5 presents and discusses the results. Section 6 concludes the paper.

2. Task Description

The GRACE shared task at IberLEF 2026 targets Argument Mining over abstracts of Randomised Controlled Trials written in Spanish. Track 1, which utilises the ClinArgES dataset [6], consists of two hierarchical subtasks that are evaluated independently.

To clearly illustrate the dataset's annotation schema and the hierarchical requirements of both subtasks, consider the complete, integrated argumentation pipeline presented in Figure 1. In this representation, the text segments form a linear chain where the empirical *Premise* directly justifies the local *Claim*, which in turn establishes the final *MajorClaim*.

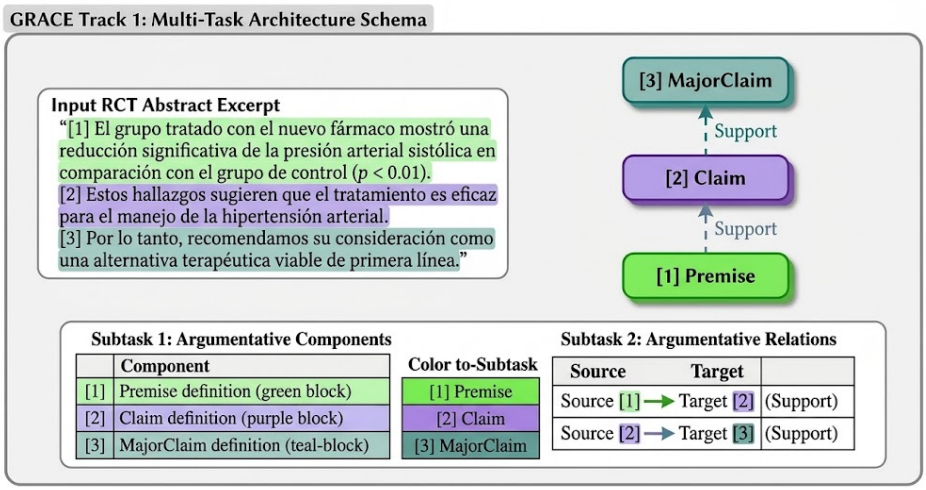

Figure 1: Hierarchical Argumentation Graph Example. The left pane illustrates character-level span boundaries (Subtask 1) while the right pane details the resolved semantic dependency tree (Subtask 2).

2.1. Subtask 1: Argumentative Component Detection

Given an RCT abstract as raw text, systems must detect and classify all argumentative components at the character level. Following the annotation scheme of Mayer et al. [2], three types of components are considered:

[1]Code available at https://github.com/lmolino03/GRACE

- **Premise**: Evidence, data, or observations that support or contest a claim (e.g., statistical results, descriptions of study outcomes).
- **Claim**: Interpretive statements or conclusions drawn from the evidence.
- **MajorClaim**: The overarching conclusion or thesis of the study, typically appearing in the conclusion section of the abstract.

Components generally correspond to complete sentences, although subordinate or coordinate clauses may be annotated as independent units depending on their argumentative role.

2.2. Subtask 2: Argumentative Relation Detection

Given the predicted argumentative components from Subtask 1, systems must classify directed relations between all pairs of components. Three relation types are considered:

- **Support**: The source component provides evidence in favour of the target.
- **Attack**: The source component contradicts or undermines the target.
- **Partial-Attack**: The source component partially contradicts the target while acknowledging some validity.

2.3. Evaluation Metrics

The official evaluation metric for both subtasks is the **macro-averaged Strict F1** score. Under the strict matching criterion, a predicted component (Subtask 1) or relation (Subtask 2) is considered correct only if its character boundaries exactly match the gold standard annotations. Additionally, for Subtask 1, a complementary relaxed metric based on token-level Intersection over Union (IoU ≥ 0.8) is also reported.

Finally, the Overall score for Track 1 is computed as the harmonic mean of the Subtask 1 and Subtask 2 official scores:

$$\text{Overall} = \frac{2 \times F1_{\text{ST1}} \times F1_{\text{ST2}}}{F1_{\text{ST1}} + F1_{\text{ST2}}} \tag{1}$$

3. System Description

Our system follows a two-stage pipeline architecture with a shared encoder. Figure 2 provides an overview. First, a pre-trained BERT-based language model encodes the input text. Then, two task-specific heads—one for entity extraction and one for relation classification—operate on the shared hidden states.

We use BETO [8] (`dccuchile/bert-base-spanish-wwm-cased`) as the shared encoder, a BERT-base model pre-trained on a large Spanish corpus with whole-word masking. The model produces contextualised representations $\mathbf{H} \in \mathbb{R}^{B \times L \times d}$, where $d = 768$.

To handle abstracts exceeding the 512-token limit, we employ a **sliding window** strategy with a configurable stride (default: 128 tokens). The tokeniser produces overlapping chunks, and entities spanning multiple chunks are aligned through offset mapping. Gradient checkpointing is enabled on the encoder to reduce VRAM consumption during training.

3.1. Subtask 1: GlobalPointer with GSSDAF

3.1.1. GlobalPointer Head

For entity extraction, we adopt the GlobalPointer framework [7], which formulates span detection as a biaffine scoring problem. Given the encoder output $\mathbf{H}$, a linear projection maps each token representation into query $\mathbf{q}_i$ and key $\mathbf{k}_i$ vectors for each entity type c:

$$[\mathbf{q}_i^{(c)}; \mathbf{k}_i^{(c)}] = \mathbf{W}^{(c)} \mathbf{h}_i + \mathbf{b}^{(c)} \tag{2}$$

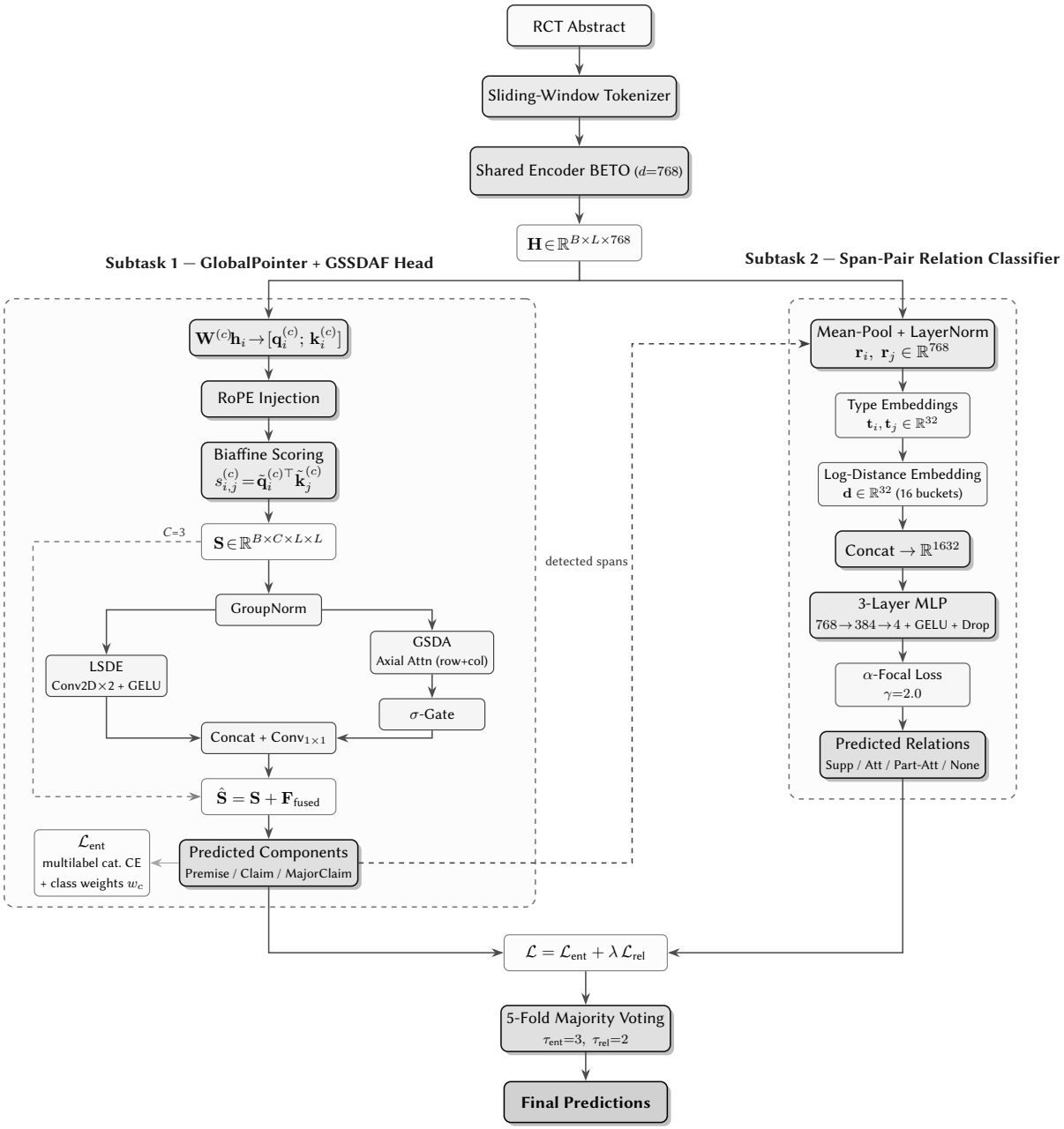

Figure 2: Full architecture of the SINAI system. A shared BETO encoder feeds two task-specific heads: (left) the **GlobalPointer + GSSDAF** head extracts argumentative components via biaffine span scoring refined with local (LSDE) and global (GSDA) span-dependency branches; (right) the **Span-Pair Relation Classifier** concatenates mean-pooled span representations with type and distance embeddings through a 3-layer MLP trained with α-Focal Loss. Dashed arrows indicate that detected entities feed the relation head and that a residual connection preserves span-logit calibration. A 5-fold majority-voting ensemble consolidates predictions at inference.

where $\mathbf{W}^{(c)} \in \mathbb{R}^{2d_h \times d}$ and d_h is the head dimension (64 in our configuration).

Rotary position embeddings (RoPE). To inject position information into the biaffine space, we apply Rotary Position Embeddings [9] to both query and key vectors. RoPE encodes absolute position through rotation matrices applied in the complex plane, enabling the dot product between queries and

keys to naturally capture relative position:

$$\text{RoPE}(\mathbf{x}, m) = \begin{pmatrix} x_1 \cos(m\theta_1) - x_2 \sin(m\theta_1) \\ x_2 \cos(m\theta_1) + x_1 \sin(m\theta_1) \\ \vdots \\ x_{d_h-1} \cos(m\theta_{d_h/2}) - x_{d_h} \sin(m\theta_{d_h/2}) \\ x_{d_h} \cos(m\theta_{d_h/2}) + x_{d_h-1} \sin(m\theta_{d_h/2}) \end{pmatrix} \tag{3}$$

where $\theta_k = 10000^{-2k/d_h}$.

The span logit matrix for entity type c is then:

$$s_{i,j}^{(c)} = \text{RoPE}(\mathbf{q}_i^{(c)}, i)^\top \text{RoPE}(\mathbf{k}_j^{(c)}, j) \tag{4}$$

producing a tensor $\mathbf{S} \in \mathbb{R}^{B \times C \times L \times L}$ where $C = |\{\text{Premise}, \text{Claim}, \text{MajorClaim}\}| = 3$.

An upper-triangular mask is applied to enforce $i \leq j$ (valid spans only), and positions outside the attention mask are set to -10^4.

3.1.2. GSSDAF Span Interaction Network

Vanilla GlobalPointer assumes independence between predicted spans, which is a poor assumption for argumentative structures where premises and claims exhibit strong co-occurrence patterns. To address this, we apply a GSSDAF interaction module [10] on top of the base logit matrix $\mathbf{S}$.

The GSSDAF module (depicted in Figure 2) processes the base logit matrix $\mathbf{S}$ through a sequence of normalisation, parallel feature extraction, and fusion steps. First, in a pre-normalisation step, a GroupNorm layer normalises $\mathbf{S}$ across the entity-type channel dimension to produce the normalised logits $\hat{\mathbf{S}}$, preventing the subsequent refinement from disrupting the calibration of the > 0 decision threshold.

Next, the normalised logits are processed through two parallel branches to model different types of span relationships. The Local Span Dependency Enhancement (LSDE) branch captures dependencies between adjacent spans (i.e., spans with similar start or end positions) by passing the normalised logits through two stacked 3×3 Conv2D layers with GELU activation:

$$\mathbf{F}_{\text{local}} = \text{Conv}_2\big(\text{GELU}(\text{Conv}_1(\hat{\mathbf{S}}))\big) \tag{5}$$

Simultaneously, the Global Span Dependency-Aware (GSDA) branch models long-range span dependencies. To avoid the prohibitive $O(L^4)$ complexity of full 2D attention, this branch applies an axial-attention mechanism [11] that reduces the complexity to $O(L^3)$ by decomposing the operation into column-wise and row-wise attention:

$$\mathbf{F}_{\text{global}} = \text{AxialAttn}_{\text{col}}(\hat{\mathbf{S}}) + \text{AxialAttn}_{\text{row}}(\hat{\mathbf{S}}) \tag{6}$$

Here, row attention models spans that share the same start position, while column attention models spans sharing the same end position.

To prevent noisy or invalid global interactions from affecting the representations, an adaptive sigmoid gating mechanism is applied to the output of the GSDA branch:

$$\mathbf{G} = \sigma(\text{Conv}_{1\times1}(\mathbf{F}_{\text{global}})) \odot \mathbf{F}_{\text{global}} \tag{7}$$

The local features $\mathbf{F}_{\text{local}}$ and the gated global features $\mathbf{G}$ are then concatenated along the channel dimension and fused through a 1×1 convolution:

$$\mathbf{F}_{\text{fused}} = \text{Conv}_{1\times1}([\mathbf{F}_{\text{local}}; \mathbf{G}]) \tag{8}$$

Finally, a residual connection adds the fused features back to the original logit matrix to produce the refined span logit matrix:

$$\hat{\mathbf{S}}_{\text{refined}} = \mathbf{S} + \mathbf{F}_{\text{fused}} \tag{9}$$

3.1.3. Entity Loss Function

For training the entity head, we use the *multilabel categorical cross-entropy* loss proposed in the original GlobalPointer work [7]. Unlike standard binary cross-entropy, this loss function treats each row of the flattened span matrix as a multi-class problem where positive spans compete against all negative spans:

$$\mathcal{L}_{\text{ent}} = \frac{1}{C}\sum_{c=1}^{C} w_c \cdot \left[\log\left(1 + \sum_{(i,j)\in\mathcal{N}^{(c)}} e^{s_{i,j}^{(c)}}\right) + \log\left(1 + \sum_{(i,j)\in\mathcal{P}^{(c)}} e^{-s_{i,j}^{(c)}}\right)\right] \tag{10}$$

where $\mathcal{P}^{(c)}$ and $\mathcal{N}^{(c)}$ denote the sets of positive and negative spans for entity type c, and w_c are dynamic class weights computed from entity frequency in the training dataset using inverse-frequency balancing:

$$w_c = \frac{N_{\text{total}}}{C \cdot N_c} \Big/ \frac{1}{C}\sum_{c'=1}^{C} \frac{N_{\text{total}}}{C \cdot N_{c'}} \tag{11}$$

3.2. Subtask 2: Span-Pair Relation Classifier

For relation classification, we employ a dedicated span-pair classifier that takes as input the shared encoder's hidden states and the detected entity positions.

3.2.1. Pair Representation and Classification

Given two entities (e_i, e_j) with token-level positions $[s_i, e_i]$ and $[s_j, e_j]$, the feature vector $\mathbf{f}$ for the pair is constructed by concatenating three components: span representations, entity type embeddings, and a distance embedding.

First, we compute the **span representations** using the mean-pooled hidden states for each entity, normalised with LayerNorm:

$$\mathbf{r}_i = \text{LN}\left(\frac{1}{e_i - s_i + 1}\sum_{k=s_i}^{e_i} \mathbf{h}_k\right) \tag{12}$$

These representations are concatenated *asymmetrically* as $[\mathbf{r}_i; \mathbf{r}_j]$ to preserve the directionality of the relation ($e_i \rightarrow e_j$). To avoid GPU memory issues with large numbers of pairs, span representations are computed in chunks of 512.

Second, we incorporate **entity type embeddings** $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^{32}$, which are learned embeddings representing the types of both arguments (Premise: 0, Claim: 1, MajorClaim: 2).

Third, we include a learned **distance embedding** $\mathbf{d} \in \mathbb{R}^{32}$ based on the logarithmically bucketised absolute token distance between the start positions of the two entities:

$$\text{bucket}(|s_i - s_j|), \quad \text{buckets} = \{0, 1, 2, 3, 4, [5, 7], [8, 15], [16, 31], \ldots\} \tag{13}$$

with 16 distance buckets total.

The final pair representation $\mathbf{f}$ has dimensionality $2d + 2 \times 32 + 32 = 2 \times 768 + 96 = 1632$. This concatenated vector is passed through a three-layer Multi-Layer Perceptron (MLP) with GELU activations and dropout ($p = 0.3$) to obtain the relation predictions:

$$\hat{y} = \mathbf{W}_3\,\text{GELU}\big(\mathbf{W}_2\,\text{GELU}(\mathbf{W}_1\mathbf{f} + \mathbf{b}_1) + \mathbf{b}_2\big) + \mathbf{b}_3 \tag{14}$$

with hidden dimensions $[d, d/2, R]$ where $R = |\{\text{NoRelation}, \text{Support}, \text{Attack}, \text{Partial-Attack}\}| = 4$.

3.2.2. Relation Loss Function

Given the extreme class imbalance in relation labels (the vast majority of pairs have no relation), we train the relation head with α-balanced Focal Loss [12]:

$$\mathcal{L}_{\text{rel}} = -\frac{1}{N}\sum_{n=1}^{N} \alpha_{y_n}(1-p_{y_n})^{\gamma}\log(p_{y_n}) \tag{15}$$

where $\gamma = 2.0$ focuses learning on hard examples, and α_{y_n} are per-class weights.

3.3. Joint Training

The total loss for joint training is a weighted combination:

$$\mathcal{L} = \mathcal{L}_{\text{ent}} + \lambda \cdot \mathcal{L}_{\text{rel}} \tag{16}$$

where λ controls the relative importance of the relation loss (set to 1.0 in our experiments).

4. Experimental Setup

4.1. Dataset

The GRACE Track 1 dataset, derived from the ClinArgES dataset [6], consists of Spanish RCT abstracts annotated with argumentative components and relations. The organisers provided a training set and a development set. For our K-Fold experiments, we partitioned the official training dataset (350 documents) into 5 folds, where each fold uses 80% of the training documents (280 documents) for training and reserves 20% (70 documents) for validation. To leverage all available data during ensembling, the validation split for each fold is combined with the separate official development dataset (50 documents), resulting in 120 validation documents per fold. This ensembling phase was preceded by a separate hyperparameter optimisation stage, where the training and development sets remained strictly separated as a static split.

4.2. Preprocessing and Model Training

To prepare the input text for the shared encoder, the clinical trial abstracts are tokenised using the pre-trained tokeniser of the base model. We set a maximum sequence length of 512 tokens, with a sliding window stride of 128 tokens, padding to the maximum length, and retaining the offset mapping to ensure proper character-to-token alignment. Using this offset mapping, entity annotations are mapped from character offsets to token positions. The label tensor for each chunk is defined as a sparse matrix $\mathbf{Y} \in \{0,1\}^{C\times L\times L}$, where $Y_{c,i,j} = 1$ if an entity of type c spans from token i to token j. To prevent out-of-memory errors during training, these label indices are stored as sparse tuples (c, i, j), and the dense matrix is generated on-the-fly during data loading.

We optimised the model hyperparameters using Optuna [13] with a Tree-structured Parzen Estimator (TPE) sampler. We ran 30 search trials over the hyperparameter space, defining the search ranges shown in Table 1. Prior to K-fold partitioning, these trials were validated solely against the static official development split (50 documents) with models trained on the official training split (350 documents), yielding a best development macro F1 score of 0.6685. This process selected a learning rate of 8.94×10^{-5}, a batch size of 2, a gradient accumulation of 1 step, a hidden head dimension of 64, and training for 45 epochs.

Using these optimal hyperparameters, the final system was trained on an NVIDIA Ampere GPU using Slurm scheduling within the PyTorch framework [14]. We employ the AdamW optimiser [15] with a fused implementation for faster execution, alongside a learning rate schedule consisting of a linear warmup for 1 epoch followed by linear decay. To stabilise training, we apply gradient clipping with a maximum norm of 1.0. Mixed-precision training is enabled via FP16 scaling using PyTorch's

Table 1
Hyperparameter search space for Optuna optimisation.

Hyperparameter	Range/Values	Best Value
Learning rate	$[4 \times 10^{-5}, 10^{-4}]$ (log)	8.94×10^{-5}
Batch size	2 (fixed)	2
Gradient accumulation	$\{1, 2, 4\}$	1
Head dimension	$\{32, 64\}$	64
Epochs	$[20, 50]$	45

`torch.amp.GradScaler`. We employ early stopping with a patience of 10 epochs based on validation loss, and implement a 5-fold cross-validation scheme using the pre-defined split configurations.

4.3. Ensemble Strategy

At inference time, we employ a majority voting ensemble over the $K = 5$ fold models:

Algorithm 1 Majority Voting Ensemble for Entities and Relations

Require: Models $\{M_1, \ldots, M_K\}$, test data $\mathcal{D}$, entity threshold τ_{ent}, relation threshold τ_{rel}

for each document $d \in \mathcal{D}$ **do**
▷ **Phase 1: Entity Ensemble**
  span_votes $\leftarrow \{\}$
  for each model M_k **do**
    entities$_k \leftarrow$ predict_entities(M_k, d)
    for each $(type, start, end) \in$ entities$_k$ **do**
      span_votes$[(type, start, end)]$ += 1
    end for
  end for
  $\mathcal{E}_{\text{final}} \leftarrow \{s \mid \text{span\_votes}[s] \geq \tau_{\text{ent}}\}$
▷ **Phase 2: Relation Ensemble**
  rel_votes $\leftarrow \{\}$
  for each pair $(e_i, e_j) \in \mathcal{E}_{\text{final}} \times \mathcal{E}_{\text{final}}$ where $i \neq j$ **do**
    for each model M_k **do**
      $r \leftarrow$ predict_relation(M_k, e_i, e_j, d)
      if $r \neq$ NoRelation **then**
        rel_votes$[(e_i, e_j, r)]$ += 1
      end if
    end for
  end for
  $\mathcal{R}_{\text{final}} \leftarrow \{(e_i, e_j, r) \mid \text{rel\_votes}[(e_i, e_j, r)] \geq \tau_{\text{rel}}\}$
end for

We set the entity vote threshold $\tau_{\text{ent}} = 3$ (majority of 5 folds) and the relation vote threshold $\tau_{\text{rel}} = 2$.

5. Results

Table 2 presents our official results on the GRACE Track 1 test set. Our system **BETO-GSSDAF-GP-KF5** (Ensemble) ranked 1st in the Track 1 leaderboard with an Overall score of 26.42. To analyse the impact of ensembling, we also report two single-model runs of the same architecture (without K-Fold ensembling) trained with different random seeds.

Our span-extraction approach is effective for detecting argumentative components in Subtask 1, achieving 36.87 strict macro F1. This formulation avoids the cascading boundary errors that plague

Table 2
Official results on the GRACE Track 1 test set. All scores are macro-averaged Strict/Relaxed F1 ($\times 100$). The last two configurations correspond to single-model runs without K-Fold ensembling, trained with different random seeds.

System / Configuration	Overall	Subtask 1 F1		Subtask 2 F1	
		Strict	Relaxed	Strict	Relaxed
BETO-GSSDAF-GP-KF5	**26.42**	**36.87**	**52.09**	15.97	24.42
BETO-GSSDAF-GP (Single, Seed 1)	23.04	30.53	44.06	15.56	23.06
BETO-GSSDAF-GP (Single, Seed 2)	22.63	29.21	43.55	**16.05**	**24.28**

sequence-labelling methods, as the model directly predicts (start, end) pairs for each entity type. Furthermore, the GSSDAF module contributes to this performance by allowing the model to reason about span co-occurrences. For instance, the presence of a Premise span in the results section of an abstract often co-occurs with a Claim span in the discussion, and the local convolutional branch captures these adjacency patterns in the span matrix.

Conversely, Subtask 2 proved considerably more challenging, yielding 15.97 strict macro F1. This is consistent with the general difficulty of relation classification in AM, which compounds the errors from entity detection (incorrect entity boundaries propagate to incorrect relation boundaries under strict evaluation) and faces severe class imbalance: the vast majority of entity pairs have no argumentative relation. Although the α-balanced Focal Loss helps mitigate this imbalance by down-weighting easy negative examples, the limited number of positive relation examples in the training data remains a bottleneck.

Our error analysis reveals several key challenges our system faces. First, boundary precision under strict evaluation represents a significant source of false negatives, where minor boundary discrepancies (such as trailing punctuation or leading whitespace) result in missed matches despite correct type predictions. Second, detecting MajorClaims remains difficult due to their scarcity in the training set, although incorporating position-based features could mitigate this. Third, relation sparsity introduces a severe class imbalance, occasionally causing the classifier to predict relations between spatially distant but argumentatively unrelated spans. Finally, pipeline error propagation is present, where missed or incorrectly boundary-aligned spans in Subtask 1 inevitably propagate to relation errors in Subtask 2.

6. Conclusions

We presented the SINAI system for the GRACE 2026 Track 1 shared task on Argument Mining in Spanish clinical trial abstracts. Our approach combines a GlobalPointer span extraction head enhanced with GSSDAF inter-span dependency modelling, a span-pair relation classifier with rich feature representations, and a 5-fold cross-validated ensemble with majority voting.

The system achieved highly competitive results, ranking 1st on the Track 1 leaderboard with an Overall score of 26.42, demonstrating the effectiveness of span-extraction methods for argumentative component detection in the biomedical domain.

Acknowledgments

We thank the organisers of the GRACE shared task at IberLEF 2026 for providing the dataset and evaluation framework. This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme,

Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP-PIDI-2024-00852) funded by the European Union under the Smart Specialization Strategy Program for Sustainability in Andalusia (S4Andalucía) 2021–2027

Declaration on Generative AI

During the preparation of this work, the authors used generative AI in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2020) 765–818.
[2] T. Mayer, S. Marro, E. Cabrio, S. Villata, Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials, Artificial Intelligence in Medicine 118 (2021) 102098.
[3] I. De la Iglesia, A. Atutxa, A. Barrena, K. Gojenola, R. Martínez, S. Montalvo, M. A. Rodríguez, S. Zakhir Puig, V. Gómez Martínez, J. Ramirez-Romero, J. M. Villa-Gonzalez, Overview of GRACE at IberLEF 2026: Granular Recognition of Argumentative Clinical Evidence, Procesamiento del Lenguaje Natural 77 (2026).
[4] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), CEUR-WS.org, 2026.
[5] A. Yeginbergen, M. Oronoz, R. Agerri, Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11687–11699.
[6] S. Zakhir-Puig, V. Gómez-Martínez, M. Á. Rodríguez-García, S. Montalvo, RaquelMartínez, ClinArgES: Clinical argumentation spanish corpus, in: Computational Models of Argument - Proceedings of COMMA 2026, Frontiers in Artificial Intelligence and Applications, IOS Press, 2026.
[7] J. Su, Globalpointer: A novel approach to named entity recognition using globally normalised biaffine spans, Blog post: https://kexue.fm/archives/8373, 2022.
[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: Proceedings of PML4DC at ICLR 2020, 2020.
[9] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, Neurocomputing 568 (2024) 127063.
[10] Y. Sun, X. Wang, H. Wu, M. Hu, Global span semantic dependency awareness and filtering network for nested named entity recognition, Neurocomputing 617 (2025) 129035. URL: https://www.sciencedirect.com/science/article/pii/S092523122401806X. doi:https://doi.org/10.1016/j.neucom.2024.129035.
[11] J. Ho, N. Kalchbrenner, D. Weissenborn, T. Salimans, Axial attention in multidimensional transformers, arXiv preprint arXiv:1912.12180 (2019).
[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of ICCV 2017, 2017, pp. 2980–2988.
[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of KDD 2019, 2019, pp. 2623–2631.
[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019.

[15] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.
[16] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: Proceedings of ECAI 2020, 2020, pp. 2108–2115.
[17] N. Green, Towards creation of a corpus for argumentation mining the biomedical genetics research literature, in: Proceedings of the First Workshop on Argumentation Mining, ACL, 2014, pp. 11–18.
[18] T. Alhindi, V. Petridis, S. Muresan, Argument mining with large language models: Opportunities and challenges, in: Proceedings of the 11th Workshop on Argument Mining, 2024.
[19] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: Proceedings of NAACL-HLT 2021, 2021, pp. 50–61.