Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Entraînement informé par solveur pour l'intégration de contraintes logiques dans l'extraction de relations d'événements

Baptiste Brunet de la Charie[1] Abdallah Arioua[1], Előd Egyed-Zsigmond[2], Thomas Veran[1]
(1) Relyens, 18 rue Edouard Rochet, 69372 Lyon, France
(2) INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, Universite de Lyon, 69621 Villeurbanne, France
baptiste.brunetdelacharie@relyens.eu, abdallah.arioua@relyens.eu,
elod.egyed-zsigmond@insa-lyon.fr, thomas.veran@relyens.eu

RÉSUMÉ

L'extraction de relations d'événements (ERE) est une tâche cruciale dans le traitement du langage naturel, impliquant l'identification et la classification des relations sémantiques entre les événements décrits dans des documents textuels. Malgré les avancées récentes grâce aux approches d'extraction conjointe, les modèles actuels rencontrent encore des défis importants, notamment une précision insuffisante dans l'extraction des relations sous-représentées mais essentielles (telles que la causalité) et d'importantes incohérences logiques parmi les relations prédites. Pour remédier à ces limitations, nous proposons un framework pour l'ERE, conçu explicitement pour améliorer la performance d'extraction et assurer la cohérence logique globale. Notre approche combine l'encodage de documents basé sur les transformateurs avec un solveur de contraintes logiques dédié qui corrige systématiquement les prédictions brutes pour garantir la cohérence dans toutes les relations d'événements extraites. Nous introduisons le concept de relations fondamentales, un sous-ensemble de relations essentielles pour préserver la cohérence logique, et nous utilisons une stratégie d'entraînement consciente du solveur afin de prioriser explicitement ces relations. Des expérimentations approfondies sur l'ensemble de données complet MAVEN-ERE démontrent que notre framework obtient pas une précision d'extraction supérieures par rapport aux méthodes d'extraction conjointe existantes.

ABSTRACT

Solver-Aware Training for Logical Constraint Integration in Event Relation Extraction

Event Relation Extraction (ERE) is a crucial task in natural language processing, involving the identification and classification of semantic relations between events described in textual content. Despite recent advancements using joint extraction approaches, current models still face substantial challenges, notably insufficient accuracy in extracting underrepresented but essential relations (such as causality) and significant logical inconsistencies among predicted relations. To address these limitations, we propose an integrated ERE framework explicitly designed to enhance extraction performance and enforce global logical consistency. Our approach combines transformer-based document encoding with a dedicated logical constraint solver that systematically corrects raw predictions to ensure consistency across all extracted event relations. We introduce the concept of *fundamental relations*, a subset of relations critical for preserving logical consistency, and utilize a solver-aware training strategy to prioritize these relations explicitly. Extensive experimentation on the comprehensive MAVEN-ERE dataset demonstrates that our framework does not achieves superior extraction accuracy compared to existing joint extraction methods.

Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA), pages 107–125.

MOTS-CLÉS : Extraction d'evenements, TALN, contraintes logiques, MAVEN-ERE, solveur.

KEYWORDS: Event Extraction, NLP, logical constraints, MAVEN-ERE, solver.

ARTICLE : **Soumis à** CORIA-TALN 2025 (CORIA).

1 Introduction

L'extraction de relations d'événements (Event Relation Extraction, ERE) est une tâche fondamentale en traitement automatique du langage naturel (natural language processing, NLP) qui consiste à identifier et catégoriser les relations sémantiques entre les événements décrits dans un contenu textuel. Son importance s'étend à divers domaines pratiques tels que l'analyse de l'actualité, le traitement de textes juridiques, et notamment la santé, où une identification et une compréhension précises des séquences d'événements peuvent influencer directement les décisions relatives à la gestion des patients et à la compréhension de la progression des maladies.

Malgré des avancées significatives, les approches actuelles de l'ERE reposent majoritairement sur des méthodes en pipeline, qui présentent souvent des limitations importantes. L'un des problèmes majeurs est la propagation d'erreurs : les inexactitudes survenant aux premières étapes (par exemple, la détection d'événements) se répercutent sur les étapes suivantes, affectant négativement la précision de l'extraction des relations (Wang et al., 2022b). Cette propagation est particulièrement problématique dans des domaines sensibles comme la santé, où une mauvaise interprétation des séquences d'événements peut entraîner une compréhension erronée de l'historique du patient ou une mauvaise analyse des événements indésirables.

Une autre limitation importante est la négligence de la cohérence logique entre les relations d'événements prédites. Les méthodes existantes n'appliquent généralement pas de manière adéquate les contraintes logiques globales, telles que l'ordre temporel ou l'asymétrie de causalité, ce qui conduit à des prédictions localement correctes mais globalement incohérentes (Han et al., 2019).

D'autres complications découlent du déséquilibre des classes, où des relations rares mais cruciales (par exemple, la causalité) sont insuffisamment représentées, ce qui compromet leur extraction précise (Chen et al., 2022).

Pour répondre à ces défis, cet article présente un framework spécifiquement conçu pour l'extraction de relations d'événements, qui traite simultanément la cohérence logique et le déséquilibre des classes. Notre approche exploite un encodage de document basé sur les transformers combiné à un classificateur multi-couches à propagation directe (multi-layer feed-forward classifier) pour une classification initiale des paires d'événements. De manière cruciale, notre framework intègre un solveur de contraintes logiques dédié, qui applique systématiquement la cohérence globale à l'ensemble des relations d'événements prédites.

Plus précisément, nous nous concentrons sur le scénario dans lequel les déclencheurs d'événements (event triggers) ont déjà été identifiés et annotés, ce qui permet à notre cadre de fonctionner directement sur les mentions d'événements. En adoptant une stratégie d'optimisation holistique, nous intégrons explicitement des contraintes de cohérence logique telles que l'ordre temporel, l'asymétrie causale et la transitivité des coréférences, ce qui améliore de manière significative la cohérence des résultats d'extraction au détriment de leur précision.

Cependant, les évaluations expérimentales menées sur le jeu de données complet MAVEN-ERE (Wang et al., 2022a) montrent que notre framework proposé n'atteint pas des améliorations notables en termes de précision d'extraction par rapport aux approches conventionnelles. Le reste de l'article est structuré comme suit : la Section 2 passe en revue les méthodologies existantes ; la Section 3 introduit les concepts clés et les notations ; la Section 4 décrit en détail notre cadre intégré ; la Section 5 fournit des évaluations expérimentales approfondies ; et enfin, la Section 6 conclut par une discussion des résultats et des perspectives pour les recherches futures.

2 Travaux Connexes

Les avancées récentes en Extraction de Relations entre Événements (Event Relation Extraction, ERE) s'articulent principalement autour de trois méthodologies : l'apprentissage structuré, les méthodes basées sur les graphes et les modèles génératifs.

Apprentissage Structuré. Les approches d'apprentissage structuré infèrent conjointement les événements et leurs relations, atténuant ainsi la propagation d'erreurs inhérente aux systèmes en pipeline. (Han et al., 2019) ont proposé un modèle de prédiction structurée neuronale qui apprend conjointement des représentations partagées pour les événements et les relations temporelles. Leur modèle utilise une inférence structurée basée sur des contraintes de programmation linéaire en nombres entiers, ce qui permet une meilleure cohérence et précision en modélisant explicitement les interdépendances entre les relations. (Deng et al., 2023) ont ensuite amélioré la prédiction structurée avec le cadre SPEECH, qui utilise une modélisation basée sur l'énergie pour représenter les dépendances complexes entre les composants événementiels au niveau des tokens, des phrases et du document. Cette approche centrée sur l'énergie facilite des prédictions structurées cohérentes.

Méthodes Basées sur les Graphes. Une autre direction importante considère l'ERE au niveau du document comme un problème de modélisation par graphe, représentant les événements comme des nœuds et leurs relations comme des arêtes. (Liu et al., 2024) ont introduit iLDF, une méthode itérative basée sur les graphes mettant l'accent sur la direction de causalité et l'affinement progressif. Leur framework construit progressivement un graphe de causalité d'événements en exploitant les relations causales identifiées avec confiance pour mettre à jour les représentations des événements, améliorant ainsi la précision d'identification. De même, (Chen et al., 2022) ont proposé ERGO, qui représente les paires d'événements candidates comme des nœuds dans un graphe relationnel complet et utilise un Relational Graph Transformer pour modéliser explicitement les interactions d'ordre supérieur et les chaînes causales transitives. Cette approche intègre notamment le raisonnement transitif directement dans le modèle.

Modèles Génératifs. Des travaux récents formulent également l'ERE comme une tâche de question-réponse générative (Wei et al., 2024), exploitant la flexibilité et la compréhension contextuelle des large language models (LLMs). (Hu et al., 2025) ont développé LLMERE, qui transforme l'ERE en un format de question-réponse à réponses multiples. En interrogeant individuellement chaque événement sur ses relations, LLMERE réduit efficacement la complexité computationnelle. Le modèle utilise également des stratégies de partitionnement pour assurer une couverture complète des relations entre événements et génère des justifications pour étayer ses prédictions, améliorant ainsi l'interprétabilité et la cohérence logique. De même, (Chen et al., 2024) ont démontré que fournir aux LLMs des contraintes logiques explicites améliore leur capacité à maintenir la cohérence dans les prédictions relationnelles, répondant ainsi aux erreurs logiques fréquentes des modèles génératifs non contraints.

3 Preliminaires

3.1 Définitions

Dans cette section, nous formalisons la tâche d'extraction des relations entre événements (ERE) et introduisons les concepts clés qui sous-tendent notre approche. Nous adoptons les définitions et la terminologie présentées dans les paragraphes suivants.

> The Beatles first UK nationwide tour lasted from 2 February 1963 until 3 March 1963. The Beatles were fourth on an eleven-act bill headed by 16-year-old Londoner, Helen Shapiro. Other acts on the tour were the Red Price Band, The Kestrels, The Honeys (UK), Dave Allen, Kenny Lynch and Danny Williams. They were also joined briefly by Billie Davis during the latter part of the tour. The tour was organised by the Arthur Howes Agency. This was the first time that the Beatles had worked with Howes.

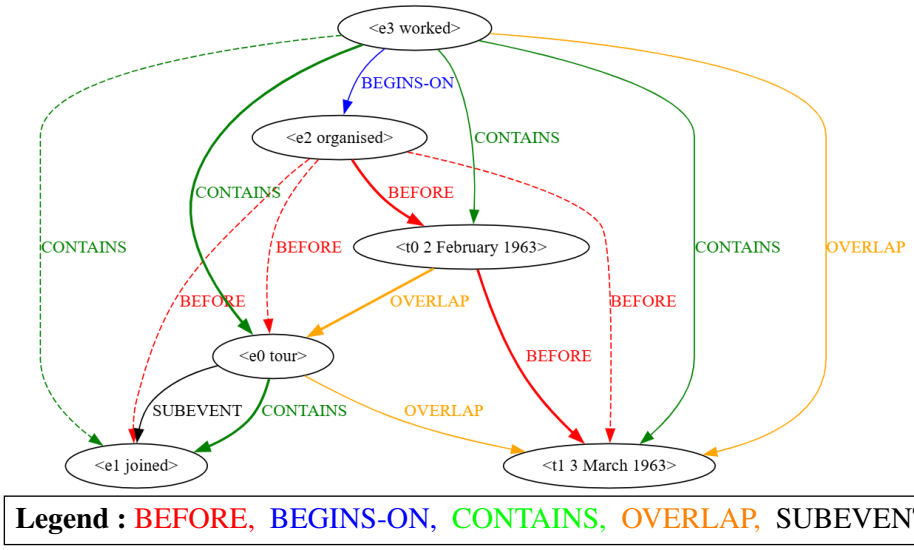


FIGURE 1 – Annotated example from the MAVEN-ERE[5.1] dataset with event mentions in blue boxes. The graph contains the color coded relations. The dashed arrows are for relations inferrable from the relations with bold arrow.

Définition d'un document Un document D est une séquence ordonnée de tokens :

$$D = \{t_1, t_2, \ldots, t_n\},$$

où n est le nombre de tokens. Nous notons $\mathcal{P}(D)$ l'ensemble des sous-ensembles (powerset) de toutes les spans de tokens dans D. Par exemple, considérons l'extrait "The Beatles first UK nationwide tour" avec les tokens $\{t_1 =$ "The", $t_2 =$ "Beatles", $t_3 =$ "first", $t_4 =$ "UK", $t_5 =$ "nationwide", $t_6 =$ "tour"$\}$. Alors $\mathcal{P}(D)$ inclut toutes les spans contiguës possibles, telles que $\{\{t_1\}, \{t_2\}, \{t_1, t_2\}, \{t_2, t_3\}, \ldots, \{t_1, t_2, t_3\}, \ldots\}$. Tout ensemble de tokens consécutifs (par exemple, $\{t_1, t_2\}$ = "The Beatles") est considéré comme une *span valide*. Remarque. Dans ce travail, un token peut représenter toute unité déterminée par le tokenizer (par exemple, une sentence-piece, une sous-unité de mot, ou un segment de texte conventionnel). Comme précisé dans les détails d'implémentation, nous adoptons finalement un tokenizer particulier pour nos expériences.

Mentions d'événements. Une mention d'événement est une span contiguë de tokens qui indique la présence ou l'occurrence d'un événement. Comme illustré dans la Figure 1

"The tour was organised by the Arthur Howes Agency."

la span "organised" est considérée comme une mention d'événement. Certains événements peuvent être mentionnés plusieurs fois dans le texte ; cependant, nous restreignons notre étude au cas où les mentions d'événements sont mutuellement exclusives — à la fois entre événements différents et au sein d'un même événement. Cette hypothèse ne constitue pas une limitation significative, car de telles chevauchements ne représentent que 0,3% des tokens de mention d'événement dans le jeu de données standard MAVEN-ERE.

Événements. Un événement dans le texte peut être décrit par une ou plusieurs mentions d'événement. Dans de nombreux cas, un événement peut n'avoir qu'une seule mention ; cependant, si plusieurs mentions d'événement situées à différents endroits du texte font référence à une même occurrence, elles forment un seul et même événement.

Expressions temporelles. Une time expression est une mention textuelle faisant référence explicitement ou implicitement à un point ou un intervalle spécifique dans le temps (par ex., "2 February 1963", "the following year", ou "two weeks later"). Bien que les time expressions diffèrent fonctionnellement des événements, elles sont généralement traitées de la même manière, comme des événements à mention unique, lors de l'extraction de relations.

Relations. Une relation capture un lien sémantique entre deux mentions d'événements (ou time expressions). Comme illustré dans la Figure 1, BEFORE(2 February 1963, 3 March 1963) indique un ordre temporel. Les types de relation sont notés R_i, avec $\mathcal{R}$ l'ensemble des types de relations. Les relations se répartissent en quatre grandes catégories :

- **Temporelles :** Spécifient l'ordre relatif et/ou le chevauchement (par ex., BEFORE, OVERLAP, BEGINS-ON) entre événements ou time expressions.
- **Causales :** Indiquent une dépendance de type cause–effet (par ex., CAUSE, PRECONDITION).
- **Sous-événement :** Déclarent qu'un événement est un sous-événement d'un autre.
- **Coréférence :** Lient deux mentions d'événements qui renvoient en réalité au même événement. Par définition, COREFERENCE est vraie entre toutes les mentions d'un même événement et fausse sinon.

Contraintes logiques et cohérence. Les contraintes logiques – notées Ψ : l'ensemble des contraintes logiques ψ_i – représentent des contraintes sémantiques qui assurent la **cohérence** entre les relations. Comme illustré dans la Figure 1, il est logiquement incohérent d'avoir simultanément BEFORE(organised, tour) et BEFORE(tour, organised) prédites comme vraies. De telles contraintes guident notre solveur pour corriger les prédictions brutes du modèle si nécessaire. Certaines de ces contraintes sont présentées dans le Tableau 1.

Relations inférables. Étant donné un ensemble de relations, une relation est considérée comme inferrable (par rapport à un ensemble de contraintes logiques Ψ) si elle peut

Nom	Formule	Implication
Asymétrie temporelle	$\text{BEFORE}(A, B) \Rightarrow \neg\text{BEFORE}(B,A)$	Évite les cycles
Transitivité temporelle	$\text{BEFORE}(A, B) \wedge \text{BEFORE}(B,C) \Rightarrow \text{BEFORE}(A,C)$	Infère un BEFORE
Transitivité causale	$\text{PRECOND}(A, B) \wedge \text{CAUSE}(B,C) \Rightarrow \text{PRECOND}(A,C)$	Infère un PRECOND
Ordre temporel	$\text{BEFORE}(A, B) \Rightarrow \neg\text{SIMULTANEOUS}(A,B)$	Préserve l'ordre
Asymétrie causale	$\text{CAUSE}(A, B) \Rightarrow \neg\text{CAUSE}(B,A)$	Causalité orientée
Symétrie de coréférence	$\text{COREF}(A, B) \Leftrightarrow \text{COREF}(B,A)$	Assure la cohérence
Transitivité de coréférence	$\text{COREF}(A, B) \wedge \text{COREF}(B,C) \Rightarrow \text{COREF}(A,C)$	Fermeture de la coréférence
Relations d'événement	$\text{REL1}(A, B) \wedge \text{COREF}(B,C) \Rightarrow \text{REL1}(A,C)$	Constance des relations

TABLE 1 – Contraintes typiques pour l'extraction de relations entre événements, où A, B et C sont des événements.

être déduite d'autres relations via des contraintes logiques. Comme illustré dans la Figure 1, BEFORE(organised, 2 February 1963) et BEFORE(2 February 1963, 3 March 1963) impliquent ensemble BEFORE(organised, 3 March 1963).

Ensemble fondamental de relations. Un fundamental relation subset $\mathcal{F}_\Psi(y)$, par rapport à un ensemble de contraintes logiques Ψ, est défini comme un ensemble maximal de relations pouvant être modifiées sans créer d'incohérences, tout en formant une base minimale à partir de laquelle toutes les autres relations peuvent être déduites. En d'autres termes, cet ensemble capture les liens essentiels, non redondants, à la fois robustes — puisqu'une modification de l'un de ses éléments ne provoque pas de violation de Ψ — et suffisants, car ils permettent d'inférer toutes les autres relations. Par exemple, dans la Figure 1, la relation

$$\text{BEFORE}(\text{organised}, \text{2 February 1963})$$

appartient à cet ensemble ; modifier sa valeur de vérité ne compromettrait pas la cohérence logique des annotations. Ainsi, tout ensemble de relations complet en termes de pouvoir inférentiel inclut nécessairement cet ensemble fondamental.

Solveur. Le solveur – noté S_Ψ, est un module de post-traitement qui ajuste les prédictions brutes — ici les logits notés $\hat{z}$ — produites par le modèle afin de garantir que toutes les relations prédites respectent l'ensemble des contraintes logiques Ψ.

Pondérations de la loss. Nous appliquons des pondérations de loss différentes pour les relations fondamentales et les relations inférables pendant l'entraînement. Les relations fondamentales reçoivent généralement des poids de loss plus élevés, car leur exactitude est cruciale pour préserver la cohérence globale et permettre l'inférence correcte des relations supplémentaires.

Ces définitions constituent la base de notre framework ERE. Dans les sections suivantes, nous décrivons comment l'encodage basé sur des transformers, l'application de contraintes de cohérence, et la correction par le solver sont intégrés pour améliorer à la fois la précision et la cohérence des relations extraites.

3.2 Formulation du problème

L'event relation extraction (ERE) vise à identifier les relations sémantiques entre les mentions d'événements au sein d'un document. La tâche ERE est formulée comme un problème de prédiction binaire multilabel et multiclass. Soit D un document représenté comme une séquence ordonnée de tokens, et soit E l'ensemble des mentions d'événements annotées dans D. Pour chaque paire ordonnée de mentions d'événements distinctes $(e_i, e_j) \in E^2$ (avec $i \neq j$) et pour chaque type de relation $r \in \mathcal{R}$, l'objectif est de prédire un label binaire indiquant si la relation r est vraie entre e_i et e_j. En dépit des progrès rapides, deux défis majeurs subsistent — en particulier dans des domaines sensibles comme la santé, où une compréhension détaillée de l'évolution d'un patient est cruciale.

(1) Prédictions incohérentes : De nombreuses approches cherchent à réduire l'incohérence des prédictions en intégrant des contraintes logiques via l'apprentissage structuré (Deng et al., 2023; Han et al., 2019). Toutefois, ces méthodes augmentent la complexité du modèle et peinent à traduire correctement les contraintes en biais inductifs. De plus, les approches reposant sur des solveurs externes nécessitent une différentiation coûteuse à travers ces solveurs, reposant sur une hypothèse d'optimalité souvent irréaliste (Tang & Khalil, 2024; Paulus et al., 2024). Il reste donc un besoin pour une solution généraliste intégrant des contraintes logiques sans rétropropagation via solveur. Notre framework répond à cette limite en découplant l'intégration des contraintes de l'apprentissage, permettant leur spécification flexible sans réentraînement, comme détaillé en Section 4, Sous-section 4.2.

(2) Performance déséquilibrée : Un défi majeur en ERE est le fort déséquilibre des classes : la majorité des paires d'événements n'a aucun label positif, et certaines relations clés, comme les relations causales, sont sous-représentées. Ce déséquilibre pousse les loss standards à privilégier les instances négatives, au détriment des positives rares mais essentielles. Pour y remédier, notre framework utilise l'Asymmetric Loss (ASL), une variante de la focal loss, qui cible spécifiquement les classes rares. Nous introduirons cette loss en Section 4, Sous-section 4.4.

4 Méthodologie

Notre approche étend un pipeline standard d'event relation extraction (ERE) en utilisant un encodeur de type transformer et un classifieur feed-forward pour générer des prédictions initiales de relations. Elle impose ensuite la cohérence logique en réajustant ces prédictions via un solveur — implémenté comme une simple multiplication matricielle entièrement différentiable — qui ajuste les logits en fonction de contraintes logiques prédéfinies. Une vue d'ensemble est présentée dans la Figure 2.

4.1 Présentation du pipeline du modèle

Extraction de caractéristiques Dans cette première étape, comme illustré dans la Figure 2, le modèle utilise un transformer-encodeur pour produire des embeddings riches et contextualisés pour chaque token du document d'entrée. Plus précisément, le document est transformé en une matrice $H \in \mathbb{R}^{|D| \times d}$, où chaque ligne correspond à la représentation sémantique d'un token dans un espace de dimension d. Cette représentation capture non seulement l'information syntaxique locale, mais

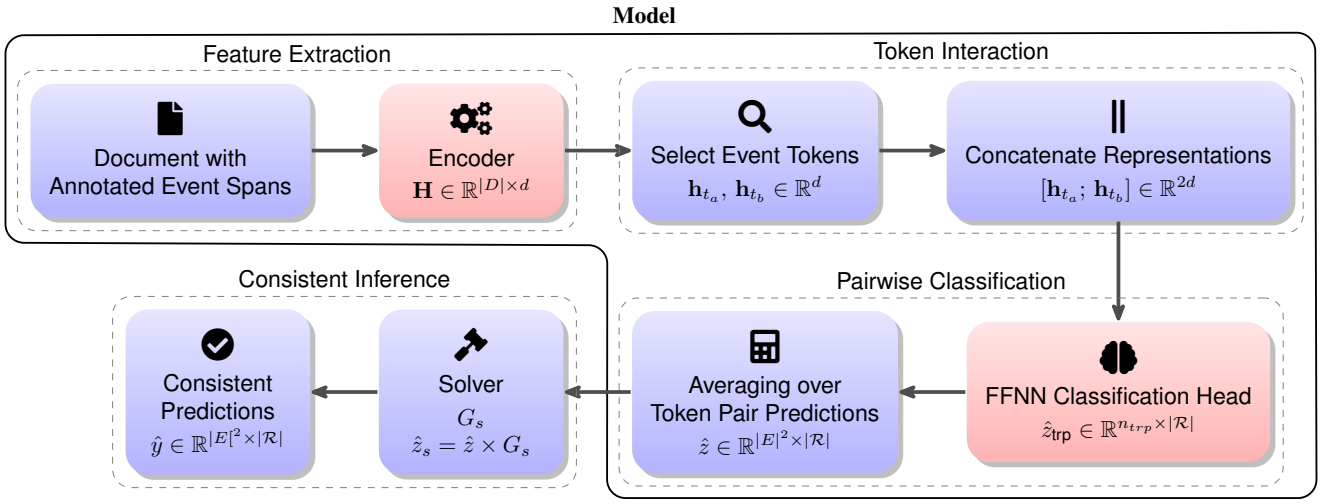


FIGURE 2 – Event Relation Extraction Model Architecture. Nodes in pink are trained. (Short Description of the pipeline).

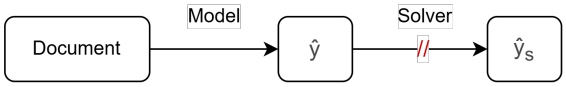
aussi les dépendances à longue portée dans le document, fournissant ainsi une base robuste pour les tâches d'ERE.

Interaction entre tokens Après l'extraction de caractéristiques, le modèle identifie les tokens associés aux mentions d'événements. Pour chaque paire d'événements candidate, tous les vecteurs représentatifs h_{ta} et h_{tb} sont extraits de la matrice d'embeddings. Ces vecteurs capturent le contexte local de chaque token de mention d'événement. Le modèle concatène ensuite ces vecteurs pour former une représentation de chaque paire de tokens $[h_{ta}; h_{tb}] \in \mathbb{R}^{2d}$, combinant ainsi efficacement les indices contextuels des deux événements.

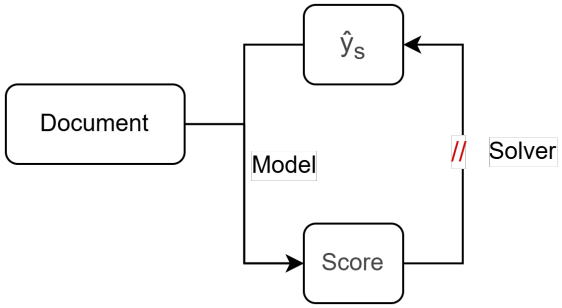
Classification paire par paire Les représentations concaténées sont passées à travers un réseau de neurones feed-forward multi-couches (FFNN), conçu pour capturer des interactions non linéaires entre les paires de tokens de mentions d'événements. Le FFNN calcule un ensemble de logits (qui sont moyennés sur les paires de tokens) $\hat{z} \in \mathbb{R}^{|E|^2 \times |\mathcal{R}|}$, où chaque logit représente une estimation non normalisée de la probabilité qu'un type de relation particulier tienne entre une paire donnée de mentions d'événements. Cette étape constitue le cœur du processus de classification, fournissant une prédiction initiale pour chaque relation d'événement possible, avant toute correction de cohérence.

Intégration du solveur de cohérence Pour résoudre les potentielles incohérences logiques dans les prédictions initiales, un solveur de cohérence est intégré au pipeline. Le solveur calcule une matrice de transformation $G_s \in \mathbb{R}^{(|E|^2|\mathcal{R}|) \times (|E|^2|\mathcal{R}|)}$ directement à partir des logits, permettant de les réajuster en fonction d'un ensemble de contraintes logiques prédéfinies. Les logits ajustés, donnés par $\hat{z}_s = G_s \times \hat{z}$, assurent une cohérence globale entre les relations prédites. Cette intégration garantit que les sorties finales respectent les règles logiques (comme l'impossibilité de relations mutuellement contradictoires), tout en restant différentiables.

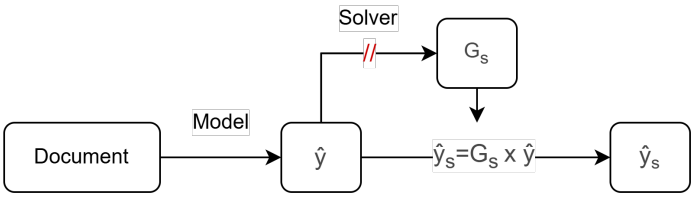
4.2 Intégration du Solveur



(a) Intégration du solveur à l'inférence seulement



(b) Intégration du solveur par optimisation du score du modèle



(c) Intégration du solveur par transformation linéaire de $\hat{y}$ par la matrice G_s

FIGURE 3 – Trois méthodes d'intégration de solveur. "//" dénote l'impossibilité de la backpropagation

Pour garantir la cohérence entre les relations prédites, nous appliquons un ensemble de contraintes basées sur des règles qui spécifient des motifs interdits. Si une contrainte est violée, la prédiction globale est considérée comme incohérente. Formellement, soit Ψ l'ensemble des contraintes, et définissons l'ensemble de toutes les évaluations cohérentes comme suit :

$$\mathcal{C}(\Psi) = \{y \mid \forall \psi \in \Psi,\ \psi(y) \text{ holds}\}. \tag{1}$$

De manière informelle, une prédiction y est cohérente si elle respecte toutes les contraintes de Ψ. Par exemple, supposons qu'une contrainte impose que si l'événement A survient avant l'événement B (noté $\text{BEFORE}(A, B)$), alors l'événement B ne peut pas survenir avant A (c'est-à-dire que $\text{BEFORE}(B, A)$ doit être faux). Si une prédiction y contient à la fois $\text{BEFORE}(A, B)$ et $\text{BEFORE}(B, A)$, elle viole cette contrainte, et donc $y \notin \mathcal{C}(\Psi)$.

Traditionnellement, un solveur $\mathcal{S}'$ est utilisé pour ajuster une évaluation initiale $\hat{y}$ afin de la rendre cohérente en résolvant :

$$\mathcal{S}'_{\Psi,\delta}(\hat{y}) = \arg \min_{\hat{y}_s \in \mathcal{C}(\Psi)} \delta(\hat{y}, \hat{y}_s), \tag{2}$$

où δ est une fonction de distance appropriée. Cette approche utilise le solveur uniquement lors de l'inférence, comme illustré dans la Figure 3a.

Une autre méthodologie a été présentée dans des travaux précédents (Han et al., 2019; Zaratiana et al., 2024), où un solveur est utilisé pour sélectionner les prédictions correctes en fonction des scores produits par le modèle, comme montré dans la Figure 3b. Dans ces approches, le modèle est entraîné à l'aide d'une loss à marge qui impose que le score de label correct dépasse celui de tout label incorrect d'une marge prédéfinie. Ainsi, en apprenant à distinguer entre les prédictions corrigées par le solveur et les labels de vérité terrain, le modèle affine davantage son mécanisme de scoring :

$$\mathcal{S}''_{\Psi,\delta}(\hat{x}) = \arg \min_{\hat{y}_s \in \mathcal{C}(\Psi)} P_{score}(\hat{y}_s, x), \tag{3}$$

où x est le document d'entrée avec les mentions d'événements, et P_{score} est le modèle de scoring évaluant la correspondance entre l'entrée x et les prédictions $\hat{y_s}$. Cela permet d'entraîner le modèle à distinguer, en termes de score, entre les annotations de vérité terrain et les prédictions corrigées par le solveur, incorporant ainsi un retour du solveur dans l'entraînement du modèle.

Cependant, cette approche néglige l'importance variable des prédictions : une erreur sur un ancrage critique peut entraîner une cascade d'erreurs, tandis qu'une erreur sur une prédiction secondaire a peu d'impact. Le feedback sur les scores ne capture pas ces dépendances structurelles.

Pour répondre à cette limitation, notre approche redéfinit le rôle du solveur de deux manières fondamentales (comme illustré dans la Figure 3c) :

1. **Sélection plutôt que modification directe :** Plutôt que de modifier toute la sortie, le solveur sélectionne un sous-ensemble minimal de prédictions (predicted fundamental) suffisant pour inférer le reste via les contraintes. Une heuristique basée sur la confiance (valeurs absolues des logits) guide une recherche gloutonne pour limiter efficacement l'espace de recherche.
2. **Matrice de transformation encodant les dépendances :** Nous calculons une Generalized Projection Matrix G_s à partir des logits $\hat{z}$ du modèle de classification par paires. La matrice G_s est déterminée en résolvant :

$$\mathcal{S}_{\Psi,\delta}(\hat{z}) = \arg \min_{G_s} \delta(\hat{z}, G_s \times \hat{z}), \tag{4}$$

sous la contrainte que la sortie binarisée appartienne à $\mathcal{C}(\Psi)$,

$$\mathbb{1}\{(\sigma(G_s) \times \hat{z}) > \eta\} \in \mathcal{C}(\Psi), \tag{5}$$

Ici, η est un seuil fixe de binarisation (avec $\hat{y} = \mathbb{1}\{\hat{p} > \eta\}$, et $\hat{p} = \sigma(\hat{z})$), et σ désigne la fonction sigmoïde. Les entrées de G_s sont comprises dans l'intervalle $[-1, 1]$, et la somme des valeurs absolues de chaque ligne est égale à 1. Par exemple, si la relation $\text{BEFORE}(A, B)$ est sélectionnée comme fondamentale, alors par la contrainte logique d'anti-symétrie temporelle, la relation $\text{NOT BEFORE}(B, A)$ est inférée comme :

$$(\hat{z}_s)^{\text{BEFORE}}_{B,A} = -1 \times (\hat{z}_s)^{\text{BEFORE}}_{A,B}.$$

Lorsqu'il existe plusieurs dépendances pour une relation donnée, notre implémentation sélectionne, par simplicité, la dépendance ayant le score absolu le plus faible pour l'utiliser dans la Generalized Projection Matrix.

L'étape globale du solveur est alors implémentée par une simple multiplication matricielle :

$$\hat{z}_s = G_s \times \hat{z}, \tag{6}$$

ce qui permet de préserver la différentiabilité du pipeline sans avoir recours à des méthodes coûteuses de différentiation boîte noire.

4.3 Fundamental Loss Weight

Notre configuration présente une redondance asymétrique : un sous-ensemble restreint de relations, appelé fundamental, suffit à inférer toutes les autres. Une relation est dite fundamental si son label n'est pas entièrement déterminé par les contraintes logiques, c'est-à-dire que changer sa valeur ne viole aucune contrainte. À l'inverse, les relations non-fondamentales ont une seule valeur cohérente, déductible des relations fondamentales.

Pour formaliser ce concept, soit y le vecteur binaire de label pour toutes les relations candidates (avec une entrée pour chaque paire ordonnée d'événements et type de relation). Pour toute relation candidate indexée par k, nous définissons l'opération suivante :

$$\mathcal{A}(y, k) = y[k \to 1 - y_k], \tag{7}$$

qui produit un nouveau vecteur de label en inversant la prédiction à l'indice k.

En utilisant cet opérateur, nous définissons l'ensemble des indices correspondant aux relations fondamentales comme :

$$\mathcal{F}_\Psi(y) = \{k \in E^2 \times |\mathcal{R}| \mid \mathcal{A}(y, k) \in \mathcal{C}(\Psi)\}. \tag{8}$$

Ainsi, une relation est fondamentale si changer son label prédit ne viole aucune contrainte. En revanche, pour une relation non-fondamentale, les contraintes logiques imposent un labrlcohérent unique, rendant sa valeur inférable à partir de l'ensemble fondamental.

Les évaluations de base révèlent que le rappel pour le sous-ensemble de relations fondamentales (par exemple, 34.1%) est inférieur au rappel global (par exemple, 58.0%). Cela indique que le modèle a des difficultés à capturer les informations non redondantes et à fort impact nécessaires à une enforcement efficace de la cohérence.

Pour atténuer ce problème, nous introduisons un Fundamental Loss Weight en amplifiant la loss pour les prédictions sur les relations fondamentales par un coefficient $\alpha \geq 1$. Soit p la probabilité prédite par le modèle et y le label de vérité terrain pour chaque relation candidate, avec $L(p, y)$ la fonction de loss (par exemple, l'entropie croisée ou une autre loss appropriée). La loss modifiée est alors définie comme suit :

$$L_{\mathcal{F}_\Psi}(p, y) = \beta \sum_{k \in \mathcal{F}_\Psi(y)} L(p_k, y_k) + \sum_{k \notin \mathcal{F}_\Psi(y)} L(p_k, y_k). \tag{9}$$

Cette loss pondérée est appliquée à la sortie du modèle de classification par paires avant l'étape du solveur, garantissant que les prédictions soient suffisamment informatives pour l'enforcement de la cohérence. En mettant davantage l'accent sur les relations fondamentales — celles qui déclenchent de multiples inférences logiques — le modèle est encouragé à réduire les erreurs sur les prédictions les plus critiques, améliorant ainsi la qualité globale de l'extraction lorsque la cohérence est appliquée.

4.4 Asymmetric Loss (ASL)

La classification multi-label souffre souvent d'un déséquilibre important, avec très peu de labels positifs par rapport aux nombreuses négatives. Les fonctions de loss conventionnelles, comme la cross-entropy, ont tendance à accorder trop d'importance aux lables abondantes ou négatives, ce qui conduit à un apprentissage insuffisant à partir des exemples positifs rares. Pour remédier à cela, des fonctions telles que la Focal Loss (FL) ont été proposées (Lin et al., 2020; Chen et al., 2022) :

$$\mathcal{L}_{\text{FL}}(y,p) = \sum_{k=1}^{K} \left[y_k \alpha (1-p_k)^{\gamma} \log(p_k) + (1-y_k)(1-\alpha)(p_k)^{\gamma} \log(1-p_k) \right], \tag{10}$$

pour un problème comportant K échantillons où $y_k \in \{0,1\}$ désigne la vérité terrain pour le $k^{\text{ième}}$ échantillon et $p_k = \sigma(\hat{z}_k)$.

Cependant, cette loss ne résout pas complètement le problème des échantillons négatifs faciles et abondants, et ne distingue pas les paramètres de focalisation entre les exemples positifs et négatifs. Comme la majorité des paires de mentions d'événements n'ont pas de relation, nous utilisons l'Asymmetric Loss (ASL) (Ridnik et al., 2021), qui repose sur deux mécanismes : asymmetric focusing et probability shifting.

L'asymmetric focusing module séparément la loss pour les échantillons positifs et négatifs en appliquant des paramètres de focalisation différents, γ_+ pour les positifs et γ_- pour les négatifs. Typiquement, on fixe $\gamma_+ = 0$, de sorte que les échantillons positifs subissent la loss cross-entropy standard, tandis qu'un γ_- plus élevé est utilisé pour réduire l'impact des négatifs faciles. En complément, le probability shifting applique un seuil fixe aux échantillons négatifs. Soit z les logits de sortie du modèle et σ la fonction sigmoïde, la probabilité de sortie du réseau $p = \sigma(\hat{z})$ est modifiée comme suit :

$$p_m = \max(p - m,\, 0),$$

où $m \geq 0$ est une marge ajustable permettant d'ignorer efficacement les négatifs très faciles (c'est-à-dire ceux pour lesquels $p < m$).

La loss pour une classification binaire unique est définie en séparant les contributions :

$$L^+ = (1-p)^{\gamma_+} \log(p), \quad \text{(pour les échantillons positifs)},$$
$$L^- = (p_m)^{\gamma_-} \log(1-p_m), \quad \text{(pour les échantillons négatifs)}.$$

Ainsi, pour un problème comportant K échantillons (chaque échantillon étant une paire ordonnée de mentions d'événements et un type de relation), la loss totale devient :

$$\mathcal{L}_{\text{ASL}}(y,p) = \sum_{k=1}^{K} \left[y_k \left(1-p_k\right)^{\gamma_+} \log(p_k) + \left(1-y_k\right) \left(p_{m,k}\right)^{\gamma_-} \log(1-p_{m,k}) \right], \tag{11}$$

Cette formulation permet à l'ASL de se concentrer sur les exemples négatifs difficiles tout en préservant la contribution des exemples positifs, ce qui permet de traiter efficacement le déséquilibre extrême entre les classes.

5 Expérimentation

Dans cette section, nous détaillons la configuration expérimentale, comparons différentes configurations de modèle et fonctions de loss, et analysons l'impact de divers hyperparamètres et choix architecturaux sur les performances.

5.1 Dataset

Le choix d'un dataset adapté pour l'extraction de relations d'événements (ERE) est crucial, notamment en santé. Il doit contenir de longs documents pour capter les dépendances étendues, annoter relations temporelles et causales, être densément annoté pour refléter la complexité réelle, et assez large pour entraîner des modèles complexes sans surapprentissage. Une bonne représentation dans la littérature est aussi souhaitable pour des comparaisons fiables.

Sélection du dataset Pour notre étude, nous avons choisi le dataset MAVEN-ERE (Wang et al., 2022a), un dataset à grande échelle qui satisfait à tous les critères mentionnés ci-dessus. MAVEN-ERE est composé de 4,480 documents en anglais issus de Wikipedia, annotés de manière exhaustive pour quatre types de relations d'événements : temporelles (par ex., BEFORE, CONTAINS, OVERLAP), causales (CAUSE, PRECONDITION), subevent et coreference. MAVEN-ERE surpasse significativement les autres datasets disponibles en termes de taille et de couverture, ce qui le rend particulièrement adapté à la modélisation des interactions complexes entre événements. D'autres datasets ont été envisagés mais écartés pour différentes raisons : (Ning et al., 2018; Cassidy et al., 2014; Glavaš et al., 2014) ne contiennent pas d'annotations causales ; (Ren et al., 2024) est limité aux textes en chinois ; et (O'Gorman et al., 2016) est trop petit en taille. En outre, le dataset EventStoryLine (Caselli & Vossen, 2017) n'a pas été retenu en raison du manque de littérature sur son utilisation pour l'extraction conjointe des relations temporelles et causales.

Description des données MAVEN-ERE fournit des annotations étendues avec un grand nombre de relations d'événements comme détaillé dans le Tableau 2. Ce cadre d'annotation dense et varié rend MAVEN-ERE particulièrement adapté pour évaluer la capacité des modèles à capturer précisément des relations complexes entre événements.

Relation Type	Number of Annotations
Temporal	1,216,217
Causal	57,992
Subevent	15,841
Coreference	103,193

TABLE 2 – Résumé des annotations de MAVEN-ERE (Wang et al., 2022a).

Découpage du dataset Nous suivons le protocole expérimental établi dans (Hu et al., 2025), en utilisant le même découpage du dataset en ensembles d'entraînement, de développement et de test selon un ratio de 70%, 10% et 20% respectivement. Ce découpage empêche tout chevauchement de documents entre les sous-ensembles, garantissant une évaluation robuste et empêchant les fuites de données.

5.2 Configuration expérimentale

Des évaluations expérimentales approfondies ont été menées selon une méthodologie systématique afin de déterminer les meilleurs réglages d'hyperparamètres pour notre approche. Les sections suivantes présentent les hyperparamètres choisis empiriquement pour maximiser la métrique de performance micro-F1, tout en assurant à la fois une efficacité d'entraînement et une cohérence à l'inférence.

Libraries Notre implémentation est réalisée en `python` à l'aide des bibliothèques `torch` (Ansel et al., 2024) et `transformers` (Wolf et al., 2020) pour les modèles, l'entraînement et l'inférence.

Hyperparamètres d'entraînement Pour toutes les expériences, nous avons utilisé un taux d'apprentissage fixe de 2×10^{-5}, et évalué la performance en termes de scores F1 micro et macro. La taille de batch était de 8 documents, ce qui a montré les meilleures performances lors d'études préliminaires, avec une variation d'environ 1% sur le score F1 micro en fonction des hyperparamètres. L'entraînement s'est systématiquement déroulé sur 20 époques, avec une évaluation du modèle sur le jeu de validation pour sélectionner le meilleur modèle.

Encoder Nous avons utilisé le Longformer de HuggingFace (`allenai/longformer-base-4096`) (Beltagy et al., 2020), avec les 6 dernières couches fine-tunées. Les expériences ont montré des rendements décroissants au-delà de ce seuil. Nous n'utilisons pas de global attention mask sur les mentions d'événements, car cela n'a pas apporté de gains significatifs. Le choix du Longformer est motivé par la longueur des textes dans le dataset MAVEN-ERE, certains dépassant les 2048 tokens, ce qui limite les encodeurs potentiels. Le modèle ModernBERT (Warner et al., 2024) a également été testé en début d'expérimentation, mais n'a pas été retenu en raison de performances significativement inférieures. Une mise à jour des connaissances ne semble pas cruciale pour cette tâche, car le domaine traité concerne des événements historiques généraux.

Classification Head La tête de classification est un réseau de neurones Feed Forward (FFNN) à 4 couches de dimensions $(2 \times h, 3 \times h, 2 \times h, |\mathcal{R}|)$, où h est la taille de la dimension cachée de l'encodeur. Chaque couche utilise un dropout de 0.1.

Fonction d'agrégation La fonction d'agrégation utilisée est la moyenne arithmétique, qui a donné de meilleures performances expérimentales que la fonction maximum et la LogSumExp (LSE).

Solveur Le solveur estime la confiance via la valeur absolue des logits et utilise une recherche gloutonne : la matrice G_s capte les dépendances inférables, en sélectionnant la relation la moins confiante en cas d'ambiguïté. La distance δ est la norme L1, avec un seuil $\eta = 0.5745$ optimisé pour un modèle ASL. Solveur et hyperparamètres restent identiques à l'entraînement et à l'inférence.

solveur pendant l'entraînement L'entraînement a été réalisé sous deux configurations différentes :

- **Sans solveur :** Le modèle est entraîné uniquement sur la loss des prédictions brutes.
- **Avec solveur (TS) :** Le modèle est entraîné avec un solveur pour imposer la cohérence logique. Le solveur est toujours utilisé à l'inférence s'il l'a été pendant l'entraînement, autrement les performances chutent.

Hyperparamètre du Fundamental Loss Weight Nous utilisons $\beta = 5$.

Pour la Focal Loss, nous utilisons $\gamma = 2$ et $\alpha = 0.25$. Pour l'Asymmetric Loss, nous utilisons $\gamma_- = 4$,

Model	Language Model	Temporal			Causal			Subevent			Coref	Average
		P	R	F1	P	R	F1	P	R	F1	F1	F1
Generation-based methods												
Llama2 (5-shot) (Wei et al., 2024)	Llama2-7b-chat	13.5	02.7	04.5	00.0	00.0	00.0	00.0	00.0	00.0	60.8	16.3
ChatGPT (4-shot) (Wei et al., 2024)	gpt-3.5-turbo	16.3	08.0	10.8	03.9	04.8	04.3	00.0	00.0	00.0	60.9	19.0
GPT4 (5-shot) (Wei et al., 2024)	gpt-4	21.6	11.6	15.1	10.4	06.0	07.6	01.9	02.4	02.1	68.4	23.3
Doc-SFT (Hu et al., 2025)	Llama2-7b-chat	35.9	18.5	24.4	25.8	30.0	27.7	19.8	27.2	23.0	82.5	39.4
LLMERE (Hu et al., 2025)	Llama2-13b-chat	50.1	**60.2**	**54.7**	**35.0**	37.2	**36.0**	26.0	30.8	28.2	**90.9**	**52.5**
Classification-based methods												
ERGO (Chen et al., 2022)	RoBERTa-base	**50.3**	52.1	51.2	31.5	25.2	28.0	26.9	18.4	21.8	89.3	47.6
Joint (Hwang et al., 2022)	RoBERTa-base	49.4	56.0	52.5	32.8	27.5	29.9	**27.3**	19.6	22.7	90.4	48.8
Split (Hwang et al., 2022)	RoBERTa-base	49.5	55.6	52.4	32.7	26.8	29.4	26.7	21.8	23.9	90.4	49.0
ProtoEM (Hu et al., 2023)	RoBERTa-base	48.9	59.9	53.8	32.5	31.3	31.8	26.2	29.7	27.9	89.8	50.8
FL[*] (Ours)	Longformer	43.1	58.4	49.6	16.0	21.8	18.5	13.3	39.4	20.0	91.8	44.8
ASL[†] (Ours)	Longformer	47.1	60.1	52.8	27.22	39.0	32.1	23.5	**43.1**	**30.4**	89.8	51.3
FW[$], IS[‡], ASL[†] (Ours)	Longformer	46.4	59.2	52.0	25.9	40.8	31.7	23.3	40.7	29.7	90.6	51.0
FW[$], IS[‡], TS[¶], ASL[†](Ours)	Longformer	46.7	59.8	52.5	26.0	**41.8**	32.1	23.8	39.7	29.7	89.4	50.9

[*]FL : Focal Loss [†]ASL : Asymmetric Loss [$]FW : Fundamental loss Weights [‡]IS : Inference Solver [¶]TS : Training Solver

TABLE 3 – Comparaisons des méthodes par classification et par génération, tiré de (Hu et al., 2025) et étendu avec nos expérimentations.

Model	Language Model	Temporal			Causal			Subevent			Global
		P	R	F1	P	R	F1	P	R	F1	F1
ASL[†] (Ours)	Longformer	48.0	62.1	54.2	32.0	18.7	23.6	34.3	**18.2**	**23.8**	52.3
FW[$], IS[‡], ASL[†] (Ours)	Longformer	**49.3**	60.3	54.2	**33.5**	**20.0**	**24.9**	33.7	16.4	22.0	52.4
FW[$], IS[‡], TS[¶], ASL[†](Ours)	Longformer	47.4	**63.7**	**54.3**	32.3	19.2	24.1	**37.0**	15.6	21.9	**52.5**

[†]ASL : Asymmetric Loss [$]FW : Fundamental loss Weights [‡]IS : Inference Solver [¶]TS : Training Solver

TABLE 4 – Comparaisons de nos expérimentations par classification sur le dataset d'évaluation pour la comparaison de micro-F1 global.

$\gamma_+ = 1$ et $m = 0.05$.

Nombre d'époques Le modèle est entraîné pendant un maximum de 10 époques, avec une évaluation de la performance sur le jeu de validation après chaque époque. Si la performance n'améliore pas la meilleure valeur enregistrée pendant trois époques consécutives, l'entraînement est interrompu prématurément et le modèle revient à l'époque ayant donné les meilleurs résultats.

5.3 Résultats et analyse

Dans cette sous-section, nous analysons les résultats issus de nos expérimentations présentés dans la Figure 3.

Impact du solveur pendant l'entraînement (TS). L'intégration du solveur a eu un impact sur la dynamique d'apprentissage : **(1) Nombre d'époques**, l'utilisation du solveur à l'entraînement et à l'inférence a nécessité environ deux fois plus d'époques (9 contre 4) pour atteindre des résultats comparables ; **(2) Complexité temporelle**, la durée de chaque époque a approximativement doublé, mettant en évidence le compromis entre l'application des contraintes logiques et l'efficacité computationnelle, principalement en raison de la nature séquentielle du solveur et des contraintes matérielles.

solveur à l'inférence (IS). L'utilisation d'un solveur à l'inférence est indépendante de l'entraînement du modèle, mais les expériences montrent une perte de performance. Cela suggère soit que le solveur utilisé produit des sorties trop éloignées de l'optimum, soit que l'évaluation cohérente la plus proche des prédictions du modèle n'est pas plus proche des labels réels.

Impact du fundamental loss weight (FW). Les fundamental loss weights n'améliorent pas les performances. Cela est attendu, car les sous-ensembles fondamentaux ne représentent qu'une faible portion des relations dans le dataset.

Fonctions de loss. Nous avons comparé deux formulations de loss pour la classification multi-label : **Focal Loss (FL)**, la loss standard pour les tâches multilabel, multiclass en présence de déséquilibres sévères dans les données ; et **Asymmetric Loss (ASL)**, qui contrebalance ce déséquilibre via un mécanisme de focalisation asymétrique et de décalage de probabilité. Comme montré dans le Tableau 3, ASL surpasse significativement FL, avec un gain de 6.5% en micro-F1.

Architectures. Nous avons également évalué plusieurs modifications architecturales : l'undersampling s'est révélé particulièrement néfaste (le micro F1 chute à 36.63%), tandis que l'ajout d'un labrlsupplémentair "no relation" a légèrement amélioré les scores malgré une complexité accrue ; par ailleurs, l'introduction d'une inconsistency loss (avec une loss supplémentaire liée au solveur et un temps d'entraînement plus long) a entraîné une baisse d'environ 1% du score F1.

Nos expériences montrent que l'Asymmetric Loss surpasse la Focal Loss, surtout pour les relations rares. La qualité de l'encodeur reste déterminante, devant la fonction de loss. Les variations architecturales apportent peu, tandis que l'undersampling dégrade les performances. Un Longformer léger surpassant Llama2-7b-chat. Bien qu'aucun gain clair ne soit observé pour notre Framework, les paramètres restent optimisables- comme la puissance du solveur, sans invalider l'approche de différentiabilité par projection linéaire. De plus, les résultats sur le dataset d'évaluation 4 montrent que le framework ne dégrade pas significativement le micro-F1 global.

6 Conclusion

Nous avons introduit un framework pour l'event relation extraction qui améliore la cohérence logique entre les relations prédites au détriment de leur précision. Ce framework apporte deux contributions principales : (1) un mécanisme modulaire basé sur un solveur qui impose des contraintes logiques globales de manière indépendante du modèle prédictif, éliminant ainsi le besoin de rétropropagation spécifique au solveur tout en maintenant une cohérence de bout en bout ; et (2) la notion de fundamental relations, un sous-ensemble minimal de relations d'événements essentiel pour inférer toutes les autres, que nous exploitons pour orienter l'entraînement informé par le solveur et concentrer l'apprentissage sur les décisions à fort impact.

Nos résultats sur le dataset MAVEN-ERE montrent que cette conception n'améliore pas les performances globales du modèle. Nous montrons que l'intégration de contraintes logiques améliore la cohérence mais compromet la précision, et que bien que notre approche surpasse des méthodes de base plus puissantes reposant sur des language models beaucoup plus volumineux, c'est la baseline qui apporte cette performance, et non le framework.

Les travaux futurs étudiront des datasets plus larges, exploreront le décodage constraint-aware pour les LLMs génératifs, et l'intégration dynamique de contraintes via l'apprentissage par renforcement.

Références

ANSEL J., YANG E., HE H., GIMELSHEIN N., JAIN A., VOZNESENSKY M., BAO B., BELL P., BERARD D., BUROVSKI E., CHAUHAN G., CHOURDIA A., CONSTABLE W., DESMAISON A., DEVITO Z., ELLISON E., FENG W., GONG J., GSCHWIND M., HIRSH B., HUANG S., KALAMBARKAR K., KIRSCH L., LAZOS M., LEZCANO M., LIANG Y., LIANG J., LU Y., LUK C. K., MAHER B., PAN Y., PUHRSCH C., RESO M., SAROUFIM M., SIRAICHI M. Y., SUK H., ZHANG S., SUO M., TILLET P., ZHAO X., WANG E., ZHOU K., ZOU R., WANG X., MATHEWS A., WEN W., CHANAN G., WU P. & CHINTALA S. (2024). PyTorch 2 : Faster machine learning through dynamic python bytecode transformation and graph compilation. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, New York, NY, USA : ACM.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer.

CASELLI T. & VOSSEN P. (2017). The Event StoryLine Corpus : A New Benchmark for Causal and Temporal Relation Extraction. In T. CASELLI, B. MILLER, M. VAN ERP, P. VOSSEN, M. PALMER, E. HOVY, T. MITAMURA & D. CASWELL, Éds., Proceedings of the Events and Stories in the News Workshop, p. 77–86, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/W17-2711.

CASSIDY T., MCDOWELL B., CHAMBERS N. & BETHARD S. (2014). An Annotation Framework for Dense Event Ordering. In K. TOUTANOVA & H. WU, Éds., Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers), p. 501–506, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-2082.

CHEN M., CAO Y., DENG K., LI M., WANG K., SHAO J. & ZHANG Y. (2022). ERGO : Event Relational Graph Transformer for Document-level Event Causality Identification. In N. CALZOLARI, C.-R. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K.-S. CHOI, P.-M. RYU, H.-H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S.-H. NA, Éds., Proceedings of the 29th International Conference on Computational Linguistics, p. 2118–2128, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

CHEN M., MA Y., SONG K., CAO Y., ZHANG Y. & LI D. (2024). Improving Large Language Models in Event Relation Logical Prediction. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 9451–9478, Bangkok, Thailand : Association for Computational Linguistics.

DENG S., MAO S., ZHANG N. & HOOI B. (2023). SPEECH : Structured prediction with energy-based event-centric hyperspheres. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 351–363, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.21.

GLAVAŠ G., ŠNAJDER J., MOENS M.-F. & KORDJAMSHIDI P. (2014). HiEve : A Corpus for Extracting Event Hierarchies from News Stories. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), p. 3678–3683, Reykjavik, Iceland : European Language Resources Association (ELRA).

HAN R., NING Q. & PENG N. (2019). Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction. 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), p. 434–444, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1041.

HU Z., LI Z., JIN X., BAI L., GUO J. & CHENG X. (2025). Large Language Model-Based Event Relation Extraction with Rationales. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., Proceedings of the 31st International Conference on Computational Linguistics, p. 7484–7496, Abu Dhabi, UAE : Association for Computational Linguistics.

HU Z., LI Z., XU D., BAI L., JIN C., JIN X., GUO J. & CHENG X. (2023). ProtoEM : A Prototype-Enhanced Matching Framework for Event Relation Extraction. DOI : 10.48550/ARXIV.2309.12892.

HWANG E., LEE J.-Y., YANG T., PATEL D., ZHANG D. & MCCALLUM A. (2022). Event-Event Relation Extraction using Probabilistic Box Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers), p. 235–244, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.26.

LIN T.-Y., GOYAL P., GIRSHICK R., HE K. & DOLLÁR P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, **42**(2), 318–327. DOI : 10.1109/TPAMI.2018.2858826.

LIU C., XIANG W. & WANG B. (2024). Identifying while learning for document event causality identification. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 3815–3827, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.210.

NING Q., WU H. & ROTH D. (2018). A multi-axis annotation scheme for event temporal relations. In I. GUREVYCH & Y. MIYAO, Éds., Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 1318–1328, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1122.

O'GORMAN T., WRIGHT-BETTNER K. & PALMER M. (2016). Richer Event Description : Integrating event coreference with temporal, causal and bridging annotation. In T. CASELLI, B. MILLER, M. VAN ERP, P. VOSSEN & D. CASWELL, Éds., Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), p. 47–56, Austin, Texas : Association for Computational Linguistics. DOI : 10.18653/v1/W16-5706.

PAULUS A., MARTIUS G. & MUSIL V. (2024). LPGD : A General Framework for Backpropagation through Embedded Optimization Layers. DOI : 10.48550/ARXIV.2407.05920.

REN Y., CAO Y., LI H., LI Y., MA Z. Z., FANG F., GUO P. & MA W. (2024). DEIE : Benchmarking Document-level Event Information Extraction with a Large-scale Chinese News Dataset. In International Conference on Language Resources and Evaluation.

RIDNIK T., BEN-BARUCH E., ZAMIR N., NOY A., FRIEDMAN I., PROTTER M. & ZELNIK-MANOR L. (2021). Asymmetric Loss For Multi-Label Classification. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), p. 82–91, Montreal, QC, Canada : IEEE. DOI : 10.1109/ICCV48922.2021.00015.

TANG B. & KHALIL E. B. (2024). PyEPO : A PyTorch-based end-to-end predict-then-optimize library for linear and integer programming. Mathematical Programming Computation, **16**(3), 297–335. DOI : 10.1007/s12532-024-00255-x.

WANG X., CHEN Y., DING N., PENG H., WANG Z., LIN Y., HAN X., HOU L., LI J., LIU Z., LI P. & ZHOU J. (2022a). MAVEN-ERE : A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, p. 926–941, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.60.

WANG Y., CHEN M., ZHOU W., CAI Y., LIANG Y., LIU D., YANG B., LIU J. & HOOI B. (2022b). Should we rely on entity mentions for relation extraction ? debiasing relation extraction with counterfactual analysis. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, p. 3071–3081, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.224.

WARNER B., CHAFFIN A., CLAVIÉ B., WELLER O., HALLSTRÖM O., TAGHADOUINI S., GALLAGHER A., BISWAS R., LADHAK F., AARSEN T., COOPER N., ADAMS G., HOWARD J. & POLI I. (2024). Smarter, Better, Faster, Longer : A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. DOI : 10.48550/arXiv.2412.13663.

WEI K., GAUTAM A. & HUANG R. (2024). Are llms good annotators for discourse-level event relation extraction ? arXiv preprint arXiv :2407.19568.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations, p. 38–45, Online : Association for Computational Linguistics.

ZARATIANA U., TOMEH N., DAUXAIS Y., HOLAT P. & CHARNOIS T. (2024). EnriCo : Enriched Representation and Globally Constrained Inference for Entity and Relation Extraction. DOI : 10.48550/arXiv.2404.12493.