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

Supervision faible pour la classification des relations discursives

Khalil Maachou[1] Chloé Braud[1,2,3] Philippe Muller[1,2,4]
(1) IRIT, Toulouse, France (2) ANITI (3) CNRS (4) Université de Toulouse prenom.nom@irit.fr

RÉSUMÉ

L'identification des relations discursives est importante pour comprendre les liens sémantiques qui structurent un texte, mais cette tâche souffre d'un manque de données qui limite les performances. D'un autre côté, de nombreux corpus discursifs existent : les divergences entre les projets d'annotation empêchent cependant de combiner directement ces jeux de données à l'entraînement. Nous proposons de résoudre ce problème en exploitant le cadre de la supervision faible, dont l'objectif est de générer des annotations à partir de sources variées, comme des heuristiques ou des modèles pré-entraînés. Ces annotations bruitées et partielles sont ensuite combinées pour entraîner un modèle sur la tâche. En combinant cette méthode avec des stratégies permettant de gérer les différences dans les jeux d'étiquettes, nous démontrons qu'il est possible d'obtenir des performances proches d'un système entièrement supervisé en s'appuyant sur une très petite partie des données d'origine, ouvrant ainsi des perspectives d'amélioration pour des domaines ou des langages à faibles ressources.

ABSTRACT

Weak supervision for discourse relation classification.

Identifying discourse relations is important for understanding the semantic links that structure a text, but this task suffers from data scarcity, which limits performance. On the other hand, many discourse corpora exist : divergences between annotation projects, however, prevent their direct combination for training puroposes. We propose to address this problem by leveraging a weak supervision framework, whose objective is to generate annotations from varied sources, such as heuristics or pre-trained models. These noisy and partial annotations are then combined to train a model on the task. By combining this method with strategies for handling differences in label sets, we demonstrate that it is possible to achieve performance close to that of a fully supervised system using a very small portion of the original data, thus opening up perspectives for improvement in low-resource domains or languages.

MOTS-CLÉS : relation de discours,supervision faible, apprentissage avec peu de données.

KEYWORDS: discourse relation, weak supervision, low-resource setting.

1 Introduction

L'annotation de données pour entraîner des modèles supervisés est un problème majeur pour l'apprentissage automatique, d'autant plus pour des tâches complexes comme l'analyse discursive automatique. Celle-ci consiste à déterminer comment les unités textuelles sont organisées par des relations sémantiques et pragmatiques, établissant par exemple des liens de cause à effet, de comparaison, de modalité ou des liens temporels. Ces informations peuvent être utilisées pour le résumé automatique (Zhang *et al.*, 2023b; Cripwell *et al.*, 2023), les systèmes questions-réponses (Prasad *et al.*, 2023; Xu *et al.*, 2023), la traduction automatique (Fernandes *et al.*, 2023; Jiang *et al.*, 2023), la détection

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 697–714.

d'émotions (Juhng *et al.*, 2023; Zhang *et al.*, 2023a), ou l'explicabilité de modèles (Devatine *et al.*, 2023). Malgré les progrès récents des modèles de langue pré-entraînés, ils nécessitent encore des données importantes pour être affinés (*fine-tuned*), et le manque de données sur des domaines ou des tâches spécialisées comme le discours limite les performances (Braud *et al.*, 2024).

Pour pallier ce manque, des approches récentes utilisent des techniques de supervision faible qui fournissent des annotations à partir de sources indirectes ou bruitées. Parmi ces méthodes, la supervision faible "programmatique" (cf. la présentation du domaine de Zhang *et al.*, 2022) génère des annotations automatiquement en utilisant des "règles d'annotation" (*label functions*), qui peuvent être des heuristiques, utiliser des bases de données ou des modèles entraînés sur des données différentes. Ces sources sont agrégées pour créer un "modèle d'annotation" (*label model*) qui produit des données de supervision (bruitée) qui peuvent servir à entraîner un modèle supervisé classique (que l'on appellera "modèle final" / *end model*).

Nous proposons une approche dans ce paradigme appliqué à la tâche de classification des relations de discours, en nous focalisant sur des corpus anglais afin de montrer que l'on peut réduire la dépendance à des jeux de données volumineux. Le domaine présentant une certaine variabilité dans les cadres et les jeux de relation, nous explorons aussi le transfert entre cadres théoriques différents, notamment RST (Mann & Thompson, 1988) et PDTB (Prasad *et al.*, 2014), et différent corpus qui peuvent varier dans les domaines et les pratiques d'annotation, même à l'intérieur d'un cadre théorique particulier. A notre connaissance ce travail est le premier à explorer ce paradigme pour cette tâche et avec des approches mêlant différents cadres théoriques. Nos contributions peuvent donc être résumées en :

— Nous montrons que la supervision faible utilisée pour générer des annotations synthétiques permet d'améliorer les performances dans un contexte de données limitées, en se rapprochant d'un système état de l'art avec seulement une fraction des données.
— Nous proposons des méthodes simples pour gérer les différents jeux de relations.
— Nous étudions les capacités de généralisation à des nouveaux domaines par transfert.

Ces expériences ont pour but d'améliorer la robustesse et les capacités de généralisation des modèles d'analyse discursive, et de réduire la dépendance aux données annotées manuellement [1].

2 Travaux reliés

La classification de relations discursives est l'une des étapes principales dans la création d'un analyseur discursif : elle consiste à trouver la catégorie de relation sémantico-pragmatique qui unit deux unités textuelles. Cette tâche est parfois divisée en deux : identifier des relations "explicites", lexicalisées par un connecteur (*mais*, *par conséquence*), et des relations "implicites" qui ne le sont pas. L'exemple de document en (a) montre une réalisation implicite de la relation *Explication*, et une explicite de *Résultat* (connecteur "En conséquence").

(a) [Un train a déraillé]_1. [(En conséquence) Il y a eu quatre blessés]_2. [Le conducteur roulait trop vite.]_3 → relations : Résultat(1,2), Explication(2,3)

De nombreux travaux ne considèrent que les implicites, mais la campagne d'évaluation DISRPT (Braud *et al.*, 2023) a montré que les performances sur les explicites chutent aussi pour certains genres ou domaines. Dans cette étude, nous considérons toutes les relations, quelque soit leur type, le but étant d'améliorer les performances générales sur la tâche pour construire des analyseurs complets. Cette tâche de classification souffre par ailleurs d'un manque de données annotées. Il existe

1. Code disponible à https://gitlab.irit.fr/melodi/andiamo/discourserelations/weakdis

cependant de nombreux corpus discursifs, si l'on ne se limite pas à un cadre, un domaine ou une langue spécifique. Cependant, la plupart des études se concentre sur un cadre spécifique, RST ou PDTB, sur une langue, l'anglais, voire sur un corpus, le PDTB (Miltsakaki *et al.*, 2004) pour les relations et le RST DT (Carlson *et al.*, 2002) pour l'analyse complète. Cette situation est due aux divergences entre les projets d'annotation et à la taille limitée de la plupart des autres corpus, mais elle accroît la rareté des données de manière artificielle : nous explorons ici le cadre de la supervision faible comme moyen d'exploiter les corpus de différents cadres.

La supervision faible (WS) fait partie d'un ensemble de méthodes dédiées au problème de rareté des données et se fonde sur l'annotation automatique de nouvelles instances. On peut distinguer la WS de l'apprentissage semi-supervisé : au lieu de propager des étiquettes existantes, en WS on utilise des connaissances diverses pour créer de l'annotation. On peut la rapprocher de l'apprentissage par transfert mais les connaissances transférés se font à partir de sources variés, pas uniquement d'un autre domaine ou langue. Parmi les méthodes classiques en WS on trouve : la supervision distante où des bases de connaissances ou des heuristiques sont utilisées pour étiqueter des données, e.g. (Awasthi *et al.*, 2020; Hoffmann *et al.*, 2011; Liang *et al.*, 2020; Ratner *et al.*, 2017), le crowd-sourcing où un ensemble d'annotateurs fournit un étiquetage de qualité variée, ou bien le co-training et le boosting pour combiner les sorties de différents modèles, e.g. (Balsubramani & Freund, 2015; Schapire & Freund, 2013). Toutes ces approches permettent de constituer des jeux de données synthétiques (bruités) et nous explorons la possibilité de les combiner pour construire des modèles plus robustes et moins gourmands en données. C'est ce que propose le cadre de la supervision faible programmatique (*Programmatic Weak Supervision*) (Zhang *et al.*, 2022) : des sources de décision faibles, ou règles d'annotation (*Label Functions, LF*) fournissent une étiquette sur une partie des données, puis une méthode d'aggrégation (*Label Model*), – un simple vote par majorité ou des méthodes plus complexes (Ratner *et al.*, 2019; Fu *et al.*, 2020; Varma *et al.*, 2019; Ren *et al.*, 2020) –, permet de produire un nouvel ensemble d'entraînement pour un modèle final (*End Model*), un modèle classique ou spécifiquement dédié à la gestion du bruit et des conflits entre étiquettes (Ratner *et al.*, 2017; Cachay *et al.*, 2021; Lan *et al.*, 2020; Parker & Yu, 2021). Dans cette étude, nous nous reposons sur ce paradigme pour combiner plusieurs approches de supervision faible.

La supervision faible pour le niveau discursif est un domaine ancien, mais encore sous-exploré. L'identification des relations implicites est la plus étudiée, notamment en exploitant les similarités avec les explicites. Les connecteurs ont été utilisés comme annotation bruitées des relations, depuis (Marcu & Echihabi, 2002) jusqu'à des travaux plus récents, e.g. (Kurfalı & Östling, 2021), étendus avec du pré-entraînement continu ou de l'apprentissage fondé sur des prompts (Xiang *et al.*, 2022; Zhou *et al.*, 2022; Wu *et al.*, 2023; Chan *et al.*, 2023). Nous testons, parmi d'autres stratégies, une annotation utilisant les connecteurs. Omura *et al.* (2024) proposent d'utiliser un modèle génératif pour augmenter les données pour certaines relations, en fournissant tous les exemples d'entraînement : l'amélioration d'$1,4\%$ seulement en micro-F1 confirme les difficultés des modèles génératifs sur cette tâche (Yung *et al.*, 2024a). Nous testons une règle fondée sur une relation spécifique et des systèmes de génération simples pour comparaison. Bourgonje & Demberg (2024) ont proposé une méthode fondée sur de la traduction automatique et / ou projection d'annotation, ce qui nécessite également d'utiliser des corpus assez larges, mais démontre le potentiel des annotations bruitées en cross-lingue. Le crowd-sourcing d'annotation a permis de produire deux corpus discursifs (Yung *et al.*, 2019; Pyatkin *et al.*, 2023) : étant donné le coût limité de ces annotations, nous testons leur inclusion comme source de connaissance faible. Concernant l'utilisation de données de différents cadres, Metheniti *et al.* (2024) ont proposé un apprentissage joint enrichi de méta-données sur le cadre et la langue, dans une approche entièrement supervisée. Pour d'autres tâches, des méthodes de supervision

faible via des règles expertes ont été testées pour l'attachement (Badene *et al.*, 2019), c'est-à-dire la détermination de quelles unités textuelles sont reliées, ou bien la segmentation (Gravellier *et al.*, 2021), c'est-à-dire l'identification des unités minimales à relier, et des méthodes de supervision distante ont permis d'améliorer le transfert entre domaines pour l'attachement via des tâches auxiliaires comme l'analyse de sentiment (Huber & Carenini, 2019, 2020) ou l'ordonnancement des unités textuelles (Li *et al.*, 2024). Nous nous intéressons ici uniquement à la tâche de classification des relations, encore sous-explorée dans un cadre de supervision faible.

3 Corpus de discours

Afin de faciliter ensuite la description de l'approche, nous présentons d'abord les jeux de données utilisés et leurs différences. Nous nous sommes limités à l'anglais et aux deux cadres les plus utilisés : RST et PDTB. Ces cadres diffèrent par de nombreux aspects, notamment les annotations en RST produisent une structure (arborescente) complète sur les documents, tandis que la couverture n'est que partielle ou locale dans les corpus PDTB. Il en résulte des différences importantes dans la nature des segments textuels formant les instances. De plus, les ensembles de relations sont différents entre les cadres, avec des définitions fondées sur des critères différents, mais aussi parfois pour les corpus d'un même cadre, ce qui empêche toute tentative de transfert direct entre les ensembles de données. Les corpus utilisés sont résumés dans la Table 1.

Nous testons deux directions dans l'apprentissage : (1) un apprentissage supervisé où l'on dispose à l'entraînement d'une petite quantité de données issues du même corpus que celui utilisé à l'évaluation, et (2) un apprentissage par transfert où le corpus d'évaluation n'a pas de section d'entraînement et où les genres textuels ne sont pas représentés dans les corpus disponibles. Le cadre (1) correspond à nos expériences principales, dont nous utilisons les résultats pour (2).

Notre principal corpus cible pour (1) est le RST DT (Carlson *et al.*, 2001), le plus utilisé pour la construction d'analyseurs discursifs pour lesquels l'identification du lien spécifique entre les segments est l'étape la plus difficile. Le RST DT est composé d'articles de journaux. L'ensemble initial de relations est très vaste mais a été réduit à 17 classes dans toutes les études ultérieures. Deux autres corpus RST sont utilisés pour tester le transfert entre domaines (2). Le corpus GUM (Zeldes, 2016) est composé de documents de nombreux domaines différents (e.g., journaux, interviews, manuels d'instruction, etc.). Bien qu'annoté en RST, ce corpus comporte un ensemble de relations qui a été adapté pour les besoins du projet, avec un total de 31 étiquettes à grain fin et 14 classes plus grossières. Le corpus GENTLE (Aoyama *et al.*, 2023) a été conçu pour tester le transfert entre domaines et ne dispose que d'un ensemble de test, avec les mêmes étiquettes que GUM.

Pour le cadre PDTB, nous utilisons le PDTB3 (Webber *et al.*, 2019), le corpus le plus grand et le plus largement utilisé pour la tâche de classification des relations. Contrairement au RST DT, il comporte une annotation des connecteurs mais, comme il ne fournit pas une couverture complète, il n'est pas associé à des analyseurs de discours. L'ensemble des relations est large mais la plupart des systèmes n'utilisent que les 23 relations de niveau 2. Nous utilisons également le corpus DiscoGem2 (Yung *et al.*, 2024b) comprenant uniquement des relations implicites avec les étiquettes du PDTB3. L'annotation a été faite par crowd-sourcing et correspond à la distribution des votes. Nous utilisons l'étiquette majoritaire. Pour les deux corpus PDTB, nous mettons en correspondance les relations annotées avec un label utilisé dans le RST DT (Annexe 8). Certaines relations du RST DT, 9 sur les 17 annotées, n'ont pas d'équivalent dans le PDTB3 (e.g. *Attribution*) ce qui conduit à un étiquetage synthétique partiel. Notons que l'annotation locale fondée sur le lexique du PDTB3 et la stratégie

de crowd-sourcing de DiscoGem2 permettent une annotation plus rapide et donc moins coûteuse, par rapport aux corpus RST utilisés. Cela motive également notre choix de cibler les corpus RST, en utilisant les corpus PDTB comme ensembles de données auxiliaires. DiscoGem2 a été obtenu auprès des auteurs, les autres *via* le benchmark DISRPT [2].

Les exemples ci-après illustrent certaines différences entre les corpus PDTB3 et RST DT, qui sont facilement comparables car certains documents sont communs [3]. L'exemple (b) montre un exemple commun aux corpus PDTB3 et RST DT, similaire en termes de contenu des arguments, mais annoté avec des étiquettes différentes. L'harmonisation définie en Table 5 ne suffit pas ici, les types de relation étant très distants : une relation comparative de *contraste vs* une relation peu informative sémantiquement de *disjonction*. L'exemple (c) illustre un autre cas où l'étiquette est la même (après avoir appliqué notre mise en correspondance) mais la segmentation est différente, le PDTB3 contenant en plus le segment en gras dans le second argument. Ces différences rendent difficiles la combinaison directe des corpus, c'est pourquoi nous proposons une méthodologie fondée sur de la supervision faible pour gérer ce que l'ont peut considérer comme du bruit dans les données auxiliaires.

(b) [Call it a fad.] [Or call it the wave of the future.]
Traduction : [*Appelez ça une mode.*] [*Ou appelez ça la vague du futur.*]
RST DT : *contrast* ; PDTB : explicite *expansion.disjunction* (PDTB et RST DT, wsj_0633)

(c) [who was derided as a "tool-and-die man"] [when GE brought him in **to clean up Kidder in 1987**]
Traduction : [*qui a été ridiculisé comme un «homme à tout faire»*] [*lorsque GE l'a engagé pour nettoyer Kidder en 1987*]
RST DT *temporal* ; PDTB : explicite *temporal.synchronous* (PDTB et RST DT, wsj_0604)

Corpus	Cadre	Type d'annotation	# Etiquettes	# instances train	# instances test
RST DT	RST	gold	17	16002	2155
GUM	RST	gold	31	19496	2617
GENTLE	RST	gold	31	-	2540
PDTB	PDTB	gold	23	43920	1610
DiscoGem	PDTB	crowdsourcing	27	59357	6595

TABLE 1 – Les corpus (anglais) utilisés, le nombre d'étiquettes différentes et le nombre d'instances. Le type d'annotation désigne soit "gold" : annotations adjudiquées, ou "crowdsourcing" : annotation multiples brutes.

4 Notre approche : supervision faible pour des scénarios à peu de données

Avant de décrire les scénarios, destinés à explorer les conditions d'utilisation d'une approche faiblement supervisée, nous décrivons la chaîne de traitement, classique en PWS (Zhang *et al.*, 2022) :

1. création des **règles d'annotation** (*Labels Functions*, LF) fournissant une étiquette pour un sous-ensemble de données. Ces règles peuvent conduire à une étiquette fausse (bruit), et à un étiquetage partiel car elles peuvent s'abstenir sur une instance pour limiter le bruit ;

2. https://huggingface.co/datasets/multilingual-discourse-hub/disrpt
3. Nous avons veillé à garder les exemples des tests absents des exemples des ensembles de test, PDTB ou RST DT.

2. aggrégation de ces règles pour produire un **modèle d'annotation** (*Label Model*) utilisé pour annoter automatiquement un ensemble de données. Nous utilisons un simple vote dans le cas où plusieurs règles s'appliquent à une même instance, en conservant la distribution des votes (normalisée à 1) ;
3. entraînement d'un **modèle supervisé final** (*end model*) sur ces données annotées automatiquement. Nous utilisons un modèle pré-entraîné reposant sur une architecture Transformers. Le modèle final est supervisé avec la distribution de probabilités sur les étiquettes fournie par le label model.

4.1 Règles d'annotation pour les relations discursives

Chaque LF cible des aspects spécifiques, de la simple utilisation d'indices lexicaux à l'exploitation des similitudes entre projets d'annotation. Une règle lexicale simple pourrait être par exemple de la forme "si le terme parce_que est présent dans une des deux unités textuelles à relier, alors la relation est *explication* ; sinon abstain". Une règle fondée sur un modèle externe consiste à utiliser comme annotation la prédiction de ce modèle. Par ailleurs, les LF sont construites à partir de données reposant sur le même cadre théorique que le corpus cible (on parlera ici de données intra-domaine), ou de données issues d'un autre cadre (dites hors-domaine). En combinant des LF fondées sur des modèles et sur des heuristiques, et des LF construites à partir de données de différents domaines / cadres, cette approche vise à maximiser l'utilisation des données disponibles, à réduire la dépendance à l'égard de l'annotation manuelle, et à améliorer la robustesse. La Table 2 résume les règles d'annotation construites.

LF	Corpus	Taille du corpus	Nombre d'étiquettes
LF à partir de données intra-domaine			
BERT - RST DT	RST DT	1600	17
RoBERTa - RST DT	RST DT	1600	17
BERT - Elaboration	RST DT	1600	2
AUTO - Generated LFs	RST DT	1600	17
LF à partir de données hors-domaine			
BERT - PDTB	PDTB3	4392	17
BERT - RST DT+DiscoGem2 filt.	RST DT + DiscoGem2	68483	8
BERT - DiscoGem2 unfilt.	DiscoGem2	66883	17
LEXICAL - Connectives PDTB	PDTB3	-	8

TABLE 2 – Synthèse des différentes règles d'annotation obtenues par affinage d'un modèle (BERT / RoBERTa), ou par génération automatique (AUTO) ou utilisant une information lexicale (LEXICAL). Nous indiquons le nombre d'exemples utilisés et le nombre d'étiquettes produites.

LF à partir d'un modèle affiné sur des données intra-domaine : ces LF utilisent comme annotations les prédictions fournies par un modèle affiné sur l'ensemble de données cible. La combinaison de différents modèles est similaire aux approches d'ensemble. Plus précisément, nous affinons 2 modèles sur le RST DT en sélectionnant des modèles Transformer largement utilisés sur la tâche (Braud *et al.*, 2023) – BERT et RoBERTa. Seule une petite partie des données d'entrainement disponibles est utilisée pour affiner/construire les LF, ici 10%, afin de simuler un environnement à faibles ressources.

LF à partir d'un modèle affiné sur des données hors domaine : ces LF sont obtenues en affinant

des modèles sur d'autres données, considérés comme hors domaine parce qu'elles reposent sur des cadres différents, et / ou consistent en des textes de genres / domaines éloignés.

Plus précisément, dans le scénario *supervision limitée* où la cible est le corpus RST DT, nous utilisons les corpus PDTB3 et DiscoGem2 – tous les deux ont des étiquettes dans le cadre du PDTB –, pour tester si la combinaison des données de distributions différentes permet d'améliorer efficacement les approches de la supervision limitée.

Étant donné que les étiquettes diffèrent, nous utilisons une correspondance entre relations (Table 5 en Annexe) qui n'est pas parfaite, ajoutant potentiellement du bruit. Pour le PDTB3, l'un des plus grands corpus existant, nous n'utilisons qu'une petite fraction des données d'entrainement disponibles pour construire les LF, ici 10%, pour tester l'utilisation d'un petit ensemble de données de grande qualité. Pour DiscoGem2, en revanche, nous utilisons l'ensemble du corpus d'entraînement en supposant que l'annotation par crowd-sourcing pourrait être plus facile à obtenir, afin d'évaluer le gain de l'utilisation d'un grand corpus potentiellement bruité. Nous testons deux stratégies avec DiscoGem2 :

— RST DT + DiscoGem2 filt. : nous affinons le modèle sur les mêmes 10% du RST DT utilisés avant et 100% du corpus DiscoGem2 en gardant les 8 étiquettes communes, les autres instances reçoivent l'étiquette *abstain* ;
— DiscoGem2 unfilt. : nous affinons le modèle uniquement sur les données DiscoGem2, mais en initialisant le modèle avec 17 étiquettes.

Dans le scénario spécifique de *transfert de domaine*, nous nous limitons au cadre RST et réutilisons donc les LF construites à partir des données RST DT, avec en addition, un affinage sur GUM afin de faire correspondre les étiquettes RST DT et GENTLE.

LF à partir d'un modèle affiné pour une relation spécifique : nous utilisons les prédictions d'un modèle affiné sur 10% du RST DT mais en focalisant sur une relation problématique car sur-représentée : *Elaboration* (40.33%). La règle attribue simplement l'étiquette *Elaboration* ou *abstain*. L'idée est similaire à une approche cascade où un modèle filtre d'abord certains exemples *via* un classifieur binaire pour se concentrer sur les exemples plus complexes, dans un cadre plus équilibré.

LF à partir des connecteurs discursifs : ces LF attribuent une étiquette en fonction de la présence d'un connecteur de discours spécifique associé à son étiquette majoritaire. Les connecteurs de discours (par exemple *but, while, because, however, therefore..*) sont utilisés pour lexicaliser une relation de discours et, lorsqu'on utilise une liste de marqueurs très fréquents, on peut considérer qu'ils ne sont pas très ambigus. Nous utilisons donc une liste prédéfinie de connecteurs de discours (Stede & Umbach, 1998) [4], et nous leur attribuons une relation selon un seuil de fréquence (relation majoritaire à plus de 80%), en rejetant les étiquettes correspondant à une distribution dispersée (l'étiquette majoritaire couvre moins de 15% des exemples).

LF générées automatiquement : ces LF sont générées automatiquement à partir d'une partie de l'ensemble de données annoté cible. Chaque LF est associée à une étiquette et est évaluée en fonction de sa précision. Nous avons sélectionné les LF ayant la plus grande précision pour chacune des 17 étiquettes. Nous utilisons l'implémentation proposée dans la bibliothèque Wrench (Zhang *et al.*, 2021) de la méthode décrite dans (Ratner *et al.*, 2016), utilisant les représentations construites par le modèle encodeur (ici Bert). la génération des LF requiert un échantillon de données annotées (nous utilisons les mêmes 10% du RST DT utilisés pour l'affinage de modèles).

4. Utilisation de DimLex : http://connective-lex.info/

4.2 Combinaison des données selon le problème ciblé

Nous explorons deux dimensions du problème de l'apprentissage, en fonction de la disponibilité des données étiquetées, selon que l'objectif est (1) de s'améliorer sur un corpus cible pour lequel on dispose d'une quantité de données limitée, ou (2) d'effectuer un transfert entre des domaines distants. L'approche repose sur la supervision faible à partir des règles d'annotation précédemment présentées qui sont combinées pour former un nouveau corpus d'entraînement pour le modèle final.

Apprentissage supervisé avec peu de données : dans ce scénario courant, seule une quantité limitée de données annotées est disponible. L'idée est donc de combiner une petite fraction de données étiquetées avec un ensemble plus important de données non étiquetées, annotées à l'aide de LF variées. Ici, nous testons différentes LF intra-domaines : les LF construites à partir d'un sous-ensemble des données cibles (combinaison de BERT- et RoBERTa-RST DT, considérée comme une baseline), les LF pour la relation sur-représentée dans les données (BERT - Elaboration) et les LF générées automatiquement (AUTO - Generated LFs). Nous évaluons également des LF construites sur des données hors domaine : les LF construites à partir d'un modèle affiné sur le PDTB3 (i.e. BERT - PDTB) ou sur DiscoGem2 (i.e. BERT - DiscoGem2 filt. / unfilt.), ou des connecteurs (GOLD - Connectives PDTB).

Apprentissage par transfert avec peu de données : cette approche consiste à évaluer sur un corpus (domaine cible) un modèle entraîné sur un autre (domaine source), ce qui teste ses capacités de généralisation. Cette question est cruciale pour le discours, où l'adaptation de domaine est réputée difficile. Nous nous limitons au cadre RST, pour focaliser plutôt sur les différences de domaine. Notre corpus cible est GENTLE, un corpus spécifiquement créé pour tester le transfert à des domaines rares, absents du RST DT et de GUM, et qui ne dispose pas d'un ensemble d'entraînement. Ce corpus est annoté en RST avec des étiquettes différentes du RST DT mais similaires à celles de GUM. Au lieu de nous fonder sur une correspondance a priori entre les jeux de relations, nous testons une approche automatique utilisant un sous-ensemble de GUM pour orienter le modèle vers les bonnes étiquettes : le modèle est affiné sur 10% du RST DT, puis sur 10% de GUM afin de tirer partie des deux ensembles de données, en espérant que les représentations apprises sur le RST DT permettent d'affiner ensuite un modèle plus informé sur GUM. Ici, les LF utilisées correspondent aux prédictions obtenus avec ce modèle successivement affiné sur RST DT et GUM, ainsi que les LF ayant amené la meilleure précision dans les expériences sur le RST DT, à savoir celles construites à partir de DiscoGem2.

5 Paramètres d'expérimentation

5.1 Systèmes de référence

Systèmes entièrement supervisés : Nous présentons les résultats reportés de (Braud *et al.*, 2024) dans un cadre entièrement supervisé, avec un entraînement sur le RST DT lors des tests sur ce même corpus, ou un entraînement sur GUM lors des tests sur GUM et GENTLE. Ce système est adapté de (Gessler *et al.*, 2021). Il s'agit d'un modèle fondé sur une architecture Transformer - BERT version de base - enrichi de traits.

Modèles génératifs : À titre de comparaison, nous testons deux stratégies simples basées sur des modèles génératifs, les annotations produites étant utilisées pour affiner ensuite un modèle :

— Génération de données : pour chaque étiquette du RST DT, le modèle génère 300 paires de segments textuels associées à leurs étiquettes. Nous supposons ignorer la véritable distribution

cible et demandons donc de produire un ensemble d'étiquettes uniformément réparties.
— Annotation de données : un modèle génératif est utilisé pour annoter directement les instances (paires de segments textuels) du RST DT.

Les deux méthodes sont évaluées avec deux configurations : (i) 0-shot : les modèles reçoivent uniquement les définitions des étiquettes extraites du manuel d'annotation du RST DT, (ii) k-shot : les modèles reçoivent en plus quelques exemples.

5.2 Détails d'implémentation

Lors de l'affinage d'un modèle pour construire une LF, nous avons utilisé : *bert-base-uncased* (Devlin *et al.*, 2019) et *roberta-base* (Liu *et al.*, 2019) [5]. BERT est aussi utilisé comme modèle final. Les valeurs d'hyper-paramètres sont 8 époques, des lots de taille 8, un un taux d'apprentissage de $5e-5$, l'optimiseur est AdamW. Les résultats sont reportés sur un seul essai.

Le modèle d'annotation utilise un vote, qui avait donné les meilleurs scores sur les systèmes de base. Ce modèle permet d'annoter le reste des données d'entraînement RST DT ou GUM, selon les scénarios. Lorsqu'il n'y a que 2 étiquettes à départager, ce qui est le cas de la combinaison de base (BERT / RoBERTa - RST DT) l'implémentation renvoie la première par défaut, ici BERT qui a pas ailleurs, une meilleure précision. Le modèle final reçoit ensuite des étiquettes correspondant à la distribution des votes : ici, par exemple, la décision de BERT, mais pondérée par $0,5$ quand RoBERTa donne une autre étiquette.

Pour générer des données sur la base d'un modèle génératif, nous utilisons : meta-llama/Meta-Llama-3-8B-Instruct (Touvron *et al.*, 2023). Pour la génération de données annotées, LLAMA doit générer 300 instances par relation. Pour l'annotation, LLAMA reçoit 5 paires de segments à annoter par appel. Dans tous les contextes, le modèle reçoit les définitions des 17 relations extraites du manuel d'annotation (Carlson *et al.*, 2001). Dans la configuration k-shot, le modèle reçoit en plus 5 exemples du RST DT.

6 Résultats

Modèle	Entraînement	Evaluation	Exactitude
Système entièrement supervisé (Braud *et al.*, 2024)	100% RST DT	RST DT	66.08%
Supervision limitée	10% RST DT	RST DT	63.17%
Modèle génératif	0% RST DT	RST DT	18.51%
Système entièrement supervisé (Gessler *et al.*, 2021)	100% GUM	GUM	64.12%
Supervision limitée	10% GUM	GUM	48.11%
Système entièrement supervisé (Gessler *et al.*, 2021)	100% GUM	GENTLE	56.26%
Transfert entre domaines	10% GUM	GENTLE	40.07%

TABLE 3 – Comparaison globale entre les meilleurs modèles de chaque approche. Les meilleurs modèles en supervision limitée utilisent un modèle affiné sur DiscoGem2 et soit une combinaison de modèle affiné sur le RST DT (partie haute), soit une combinaison de modèles affinés sur GUM.

Approche supervisée avec peu de données Comme le montre la Table 3, notre meilleur modèle est à

5. https://huggingface.co/

moins de 3 points derrière le système entièrement supervisé, tout en utilisant 10% des données du corpus cible. Ce système est obtenu en combinant des LF construites en utilisant 10% du RST DT avec deux modèles différents (BERT et RoBERTa) et des LF construites sur DiscoGem2, annoté par crowd-sourcing. L'amélioration observée peut être due au fait que DiscoGem2 ne comporte que de relations implicites, permettant une amélioration de la prédiction pour ce type de relations réputé plus difficile. Ces performances démontrent la capacité d'un système simple de supervision faible, fondé sur un vote, à tirer profit d'annotations diverses, partielles et potentiellement bruitées.

La précision des LF est calculée à l'étape de création des règles, elle permet de regarder la performance d'une règle simple, sans aggrégation. La règle construite à partir de BERT (57, 51%) a de meilleures performances qu'avec RoBERTa (55, 42%), la version de base de RoBERTa est moins performante pour cette tâche mais leur combinaison est bénéfique, cf Table 2. Par ailleurs, l'utilisation combinée des modèles, sur le corpus cible et sur un corpus hors-domaine, permet d'améliorer de presque 6% les performances que l'on aurait obtenues par un entraînement limité aux 10% de données disponibles (57, 51 avec BERT seul *vs* 63, 17% avec notre meilleur combinaison). Les scores par relations détaillés en Annexe 9 montre une amélioration générale, naturellement en particulier pour les relations présentes dans DiscoGem2.

Modèles de comparaison Les modèles construits comme point de comparaison sont uniquement évalués sur le RST DT. Comme le montre la Table 3, les résultats obtenus avec un modèle génératif sont très faibles, avec au mieux 18, 51% (génération en 0-shot). Nous ne pouvons pas effectuer de comparaison directe avec l'état de l'art, limité aux relations implicites sur le PDTB. Cependant, les résultats sont conformes à ceux obtenus par exemple par Omura *et al.* (2024) où l'utilisation d'un modèle GPT dans un contexte few-shot (définitions et 100 exemples) donne au mieux 29, 4 en micro-F1 contre 64, 2 avec un modèle RoBERTa large.

Apprentissage par transfert avec peu de données Ces expériences visent à évaluer la capacité de transfert entre domaines. Ici, le système de référence est entraîné sur 10% du RST DT puis 10% de GUM, cette seconde étape permettant de diriger vers le jeu d'étiquettes cible. Une fois enrichi avec notre meilleur jeu de LF déterminé par les expériences précédentes (fondé sur DiscoGem2), ce système obtient 48, 11% d'exactitude sur GUM et 40, 07% sur GENTLE. Contrairement à ce que nous avions observé précédemment, ici la combinaison des sources de supervision faible ne permet pas d'approcher un système supervisé sur la totalité des données et évalué sur GUM, on observe une perte d'environ −17% (64.12 vs 48.11). On remarque également que la perte des modèles lors du transfert de domaine, qui est attendue, est similaire entre notre approche (−8, 4, 40.07%) et le contexte entièrement supervisé (−7, 8, 56.26). Comme RST DT et GUM utilisent le même cadre sous-jacent, on pourrait s'attendre à ce qu'ils s'entraident davantage, mais ici, l'écart de domaine pourrait avoir un impact important sur les performances.

Effets des différentes LF La Table 4 contient les performances pour différentes LF et leur combinaison. La partie supérieurs ("BASE +") correspond à une évaluation sur le RST DT, la partie inférieure (BASE-T +") correspond à une évaluation sur GUM / GENTLE. Pour le RST DT, la simple combinaison des résultats de deux modèles pré-entraînés (considéré comme baseline) donne déjà une exactitude assez haute de 61, 44, dépassant les précisions de chaque règle prises individuellement. Leur combinaison avec un modèle entraîné sur un corpus provenant d'un cadre différent – ici PDTB –, conduit par contre à une baisse de performance (58, 85), indiquant les difficultés d'une telle combinaison. L'utilisation d'une règle dédiée à une relation spécifique n'aide pas, prise seule (61, 38), comme combiné au modèle utilisant DiscoGem2 (62.81 *vs* 63.17). Notre meilleur modèle repose sur la combinaison de la baseline avec DiscoGem2, un corpus qui repose également sur le

Combinaison de LF	Exactitude du modèle final
BASE = BERT-RSTDT + RoBERTa-RSTDT	61.44
BASE + AUTO-RSTDT	61.75
BASE + BERT-PDTB	58.85
BASE + BERT-elaboration	61.38
BASE + BERT-DiscoGem2+RST DT filt.	**63.17**
BASE + BERT-DiscoGem2+RST DT filt. + BERT-elaboration	62.81
BASE + BERT-DiscoGem2 unfilt.	60.33
BASE-T = BERT-GUM + RoBERTa-GUM	36.65 / 36.91
BASE-T + BERT-DiscoGem2	48.11 / **40.07**

TABLE 4 – Exactitude du modèle final pour différentes combinaisons de règles d'annotation (LF). Les systèmes incluant "BASE" sont tous évalués sur le RST DT ; les systèmes incluant "BASE-T" sont dédiés au transfert et évalué sur GUM / GENTLE.

cadre PDTB. Nous supposons qu'il y a ici un effet de taille - nous utilisons l'intégralité de DiscoGem alors que nous nous limitons à 10% de PDTB -, et peut-être une amélioration ciblée sur les relations implicites - DiscoGem ne contient que des implicites, alors que PDTB contient tous les types de relations. Cette hypothèse devrait être explorée plus avant dans de futures expériences en extrayant les relations implicites du PDTB3.

Finalement, notons que le recours à des règles créées automatiquement ("BASE + AUTO-RST DT") conduit à une légère amélioration de la baseline, ce qui est encourageant étant donné que cela limite la nécessité de créer des LF, et ouvre la voie à de futures expérimentations sur cette technique.

Pour l'adaptation de domaine, nous ne testons que les meilleurs modèles, donc une combinaison de RST DT et DiscoGem2, et également GUM ici, pour permettre l'adaptation aux différences entre les ensembles de relations. La baseline, correspondant à la combinaison de BERT et RoBERTa ajustée sur GUM, conduit à une précision de 36.65% sur GUM et 36.91% sur GENTLE, alors que les meilleurs modèles atteignent resp. 48.11 et 40.07, démontrant un effet très positif de la supervision faible, un résultat encourageant pour les travaux futurs en adaptation de domaine pour le discours.

7 Conclusion

La taille limitée des données annotées manuellement est un problème omniprésent en TAL, surtout pour les architectures neuronales. Nous avons réalisé de premières expérimentations dans un cadre de supervision faible pour une tâche sémantique de haut niveau, la classification des relations discursives, qui, en plus de la rareté des données, souffre également d'importantes divergences entre les projets d'annotation. Nous avons montré que les méthodes de supervision faible permettent d'obtenir des performances proches de l'état de l'art avec seulement 10% des données, en tirant avantage d'autres corpus. De plus, l'aggrégation de sources de supervision faible peut aider dans un contexte d'adaptation de domaine, même si ici les résultats sont très inférieurs à l'état de l'art. Nous prévoyons d'étendre nos expériences à d'autres méthodes de supervision faible programmatique, à d'autres sources d'information, et d'évaluer ces stratégies dans d'autres langues et d'autres domaines. Enfin, nous chercherons à construire un analyseur discursif complet en combinant ces méthodes.

Remerciements

Ces travaux sont partiellement financés par le projet AnDiaMO (ANR-21-CE23-0020) et l'ANR (ANR-19-PI3A-0004) à travers l'institut 3IA inter-disciplinaire ANITI, qui fait partie du programme français "Investing for the Future — PIA3". Chloé Braud et Philippe Muller font aussi partie du programme DesCartes, supporté par l'ANR, et le Prime Minister's Office de Singapour sous l'égide de son programme Campus for Research Excellence and Technological Enterprise (CREATE).

Références

AOYAMA T., BEHZAD S., GESSLER L., LEVINE L., LIN J., LIU Y. J., PENG S., ZHU Y. & ZELDES A. (2023). GENTLE : A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.law-1.17.

AWASTHI A., GHOSH S., GOYAL R. & SARAWAGI S. (2020). Learning from rules generalizing labeled exemplars. *In proceedings of ICLR*.

BADENE S., THOMPSON K., LORRÉ J.-P. & ASHER N. (2019). Weak supervision for learning discourse structure. 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. 2296–2305, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1234.

BALSUBRAMANI A. & FREUND Y. (2015). Scalable semi-supervised aggregation of classifiers. *Advances in Neural Information Processing Systems*, **28**.

BOURGONJE P. & DEMBERG V. (2024). Generalizing across languages and domains for discourse relation classification. In T. KAWAHARA, V. DEMBERG, S. ULTES, K. INOUE, S. MEHRI, D. HOWCROFT & K. KOMATANI, Éds., *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 554–565, Kyoto, Japan : Association for Computational Linguistics. DOI : 10.18653/v1/2024.sigdial-1.47.

BRAUD C., LIU Y. J., METHENITI E., MULLER P., RIVIÈRE L., RUTHERFORD A. & ZELDES A. (2023). The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, p. 1–21, Toronto, Canada : The Association for Computational Linguistics. DOI : 10.18653/v1/2023.disrpt-1.1.

BRAUD C., ZELDES A., RIVIÈRE L., LIU Y. J., MULLER P., SILEO D. & AOYAMA T. (2024). DISRPT : A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 4990–5005, Torino, Italia : ELRA and ICCL.

CACHAY S. R., BOECKING B. & DUBRAWSKI A. (2021). End-to-end weak supervision. *Advances in Neural Information Processing Systems*, **34**, 1845–1857.

CARLSON L., MARCU D. & OKUROVSKY M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

CARLSON L., OKUROWSKI M. E. & MARCU D. (2002). *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

CHAN C., LIU X., CHENG J., LI Z., SONG Y., WONG G. & SEE S. (2023). DiscoPrompt : Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 35–57, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.4.

CRIPWELL L., LEGRAND J. & GARDENT C. (2023). Context-aware document simplification. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 13190–13206, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.834.

DEVATINE N., MULLER P. & BRAUD C. (2023). An integrated approach for political bias prediction and explanation based on discursive structure. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 11196–11211, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.711.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

FERNANDES P., YIN K., LIU E., MARTINS A. & NEUBIG G. (2023). When does translation require context ? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 606–626, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.36.

FU D., CHEN M., SALA F., HOOPER S., FATAHALIAN K. & RÉ C. (2020). Fast and three-rious : Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, p. 3280–3291 : PMLR.

GESSLER L., BEHZAD S., LIU Y. J., PENG S., ZHU Y. & ZELDES A. (2021). DisCoDisCo at the DISRPT2021 shared task : A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, p. 51–62, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.disrpt-1.6.

GRAVELLIER L., HUNTER J., MULLER P., PELLEGRINI T. & FERRANÉ I. (2021). Weakly supervised discourse segmentation for multiparty oral conversations. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 1381–1392, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.104.

HOFFMANN R., ZHANG C., LING X., ZETTLEMOYER L. & WELD D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics : human language technologies*, p. 541–550.

HUBER P. & CARENINI G. (2019). Predicting discourse structure using distant supervision from sentiment. 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. 2306–2316, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1235.

HUBER P. & CARENINI G. (2020). MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7442–7457, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.603.

JIANG Y. E., LIU T., MA S., ZHANG D., SACHAN M. & COTTERELL R. (2023). Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7853–7872, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.435.

JUHNG S., MATERO M., VARADARAJAN V., EICHSTAEDT J., V GANESAN A. & SCHWARTZ H. A. (2023). Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 1500–1511, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-short.128.

KURFALI M. & ÖSTLING R. (2021). Let's be explicit about that : Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, p. 1–10.

LAN O., HUANG X., LIN B. Y., JIANG H., LIU L. & REN X. (2020). Learning to contextually aggregate multi-source supervision for sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 2134–2146.

LI C., BRAUD C., AMBLARD M. & CARENINI G. (2024). Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In M. STRUBE, C. BRAUD, C. HARDMEIER, J. J. LI, S. LOAICIGA, A. ZELDES & C. LI, Éds., *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, p. 161–176, St. Julians, Malta : Association for Computational Linguistics.

LIANG C., YU Y., JIANG H., ER S., WANG R., ZHAO T. & ZHANG C. (2020). Bond : Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, p. 1054–1064.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A robustly optimized BERT pretraining approach. *arXiv preprint arXiv :1907.11692*.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization. *Text*, **8**, 243–281.

MARCU D. & ECHIHABI A. (2002). An unsupervised approach to recognizing discourse relations. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 368–375, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073145.

METHENITI E., BRAUD C. & MULLER P. (2024). Feature-augmented model for multilingual discourse relation classification. In M. STRUBE, C. BRAUD, C. HARDMEIER, J. J. LI, S. LOAICIGA, A. ZELDES & C. LI, Éds., *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, p. 91–104, St. Julians, Malta : Association for Computational Linguistics.

MILTSAKAKI E., PRASAD R., JOSHI A. & WEBBER B. (2004). The penn discourse treebank.

OMURA K., CHENG F. & KUROHASHI S. (2024). An empirical study of synthetic data generation for implicit discourse relation recognition. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 1073–1085, Torino, Italia : ELRA and ICCL.

PARKER J. & YU S. (2021). Named entity recognition through deep representation learning and weak supervision. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3828–3839.

PRASAD A., BUI T., YOON S., DEILAMSALEHY H., DERNONCOURT F. & BANSAL M. (2023). MeetingQA : Extractive question-answering on meeting transcripts. 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. 15000–15025, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.837.

PRASAD R., WEBBER B. & JOSHI A. (2014). Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, **40**(4), 921–950. DOI : 10.1162/COLI_a_00204.

PYATKIN V., YUNG F., SCHOLMAN M. C., TSARFATY R., DAGAN I. & DEMBERG V. (2023). Design choices for crowdsourcing implicit discourse relations : Revealing the biases introduced by task design. *arXiv e-prints*, p. arXiv–2304.

RATNER A., BACH S. H., EHRENBERG H., FRIES J., WU S. & RÉ C. (2017). Snorkel : Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, p. 269 : NIH Public Access.

RATNER A., HANCOCK B., DUNNMON J., SALA F., PANDEY S. & RÉ C. (2019). Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 4763–4771.

RATNER A. J., DE SA C. M., WU S., SELSAM D. & RÉ C. (2016). Data programming : Creating large training sets, quickly. In *In proceedings of NeurIPS*.

REN W., LI Y., SU H., KARTCHNER D., MITCHELL C. & ZHANG C. (2020). Denoising multi-source weak supervision for neural text classification. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 3739–3754, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.334.

SCHAPIRE R. E. & FREUND Y. (2013). Boosting : Foundations and algorithms. *Kybernetes*, **42**(1), 164–166.

STEDE M. & UMBACH C. (1998). Dimlex : A lexicon of discourse markers for text generation and understanding. In *Proceedings of COLING*.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models.

VARMA P., SALA F., SAGAWA S., FRIES J., FU D., KHATTAR S., RAMAMOORTHY A., XIAO K., FATAHALIAN K., PRIEST J. *et al.* (2019). Multi-resolution weak supervision for sequential data. *Advances in Neural Information Processing Systems*, **32**.

WEBBER B., PRASAD R., LEE A. & JOSHI A. (2019). *The Penn Discourse TreeBank 3.0 Annotation Manual*. Rapport interne, University of Edinburgh, Interactions, LLC, University of Pennsylvania.

WU H., ZHOU H., LAN M., WU Y. & ZHANG Y. (2023). Connective prediction for implicit discourse relation recognition via knowledge distillation. 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. 5908–5923, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.325.

XIANG W., WANG Z., DAI L. & WANG B. (2022). ConnPrompt : Connective-cloze prompt learning for implicit discourse relation recognition. 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. 902–911, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

XU F., SONG Y., IYYER M. & CHOI E. (2023). A critical evaluation of evaluations for long-form question answering. 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. 3225–3245, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.181.

YUNG F., AHMAD M., SCHOLMAN M. & DEMBERG V. (2024a). Prompting implicit discourse relation annotation. In S. HENNING & M. STEDE, Éds., *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, p. 150–165, St. Julians, Malta : Association for Computational Linguistics.

YUNG F., DEMBERG V. & SCHOLMAN M. (2019). Crowdsourcing discourse relation annotations by a two-step connective insertion task. In A. FRIEDRICH, D. ZEYREK & J. HOEK, Éds., *Proceedings of the 13th Linguistic Annotation Workshop*, p. 16–25, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4003.

YUNG F., SCHOLMAN M., ZIKANOVA S. & DEMBERG V. (2024b). DiscoGeM 2.0 : A parallel corpus of English, German, French and Czech implicit discourse relations. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 4940–4956, Torino, Italia : ELRA and ICCL.

ZELDES A. (2016). The GUM corpus : Creating multilayer resources in the classroom. In *Proceedings of LREC*.

ZHANG D., CHEN F. & CHEN X. (2023a). DualGATs : Dual graph attention networks for emotion recognition in conversations. 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. 7395–7408, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.408.

ZHANG J., HSIEH C.-Y., YU Y., ZHANG C. & RATNER A. (2022). A survey on programmatic weak supervision. arXiv 2202.05433.

ZHANG J., YU Y., LI Y., WANG Y., YANG Y., YANG M. & RATNER A. (2021). WRENCH : A comprehensive benchmark for weak supervision. In *NeurIPS*.

ZHANG S., WAN D. & BANSAL M. (2023b). Extractive is not faithful : An investigation of broad unfaithfulness problems in extractive summarization. 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. 2153–2174, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.120.

ZHOU H., LAN M., WU Y., CHEN Y. & MA M. (2022). Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics : EMNLP 2022*, p. 3848–3858, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

8 Annexe : Correspondance entre les jeux de relations RST DT / PDTB

Afin d'adapter les jeux de relations entre les différents corpus, nous proposons un mapping simple, décrit dans la Table 5.

PDTB label	RST DT corresponding label
expansion.instantiation	background
contingency.cause+belief	cause
contingency.cause	cause
contingency.cause+speechact	cause
contingency.negative-cause	cause
comparison.similarity	comparison
contingency.condition+speechact	condition
contingency.negative-condition	condition
contingency.condition	condition
comparison.contrast	contrast
comparison.concession	contrast
comparison.concession+speechact	contrast
expansion.level-of-detail	elaboration
contingency.purpose	enablement
expansion.conjunction	joint
expansion.disjunction	joint
expansion.substitution	joint
expansion.exception	joint
expansion.manner	manner-means
hypophora	topic-comment
expansion.equivalence	summary
temporal.asynchronous	temporal
temporal.synchronous	temporal

TABLE 5 – Correspondance entre les étiquettes des relations dans le PDTB3 / DiscoGem2 et celles du RST DT.

9 Annexe : Scores par relation sur le RST DT

	Baseline	Best	Baseline	Best	Baseline	Best
Etiquette	Précision (%)		Rappel (%)		F-score (%)	
Attribution	78.78	82.85	85.21	88.33	81.86	**85.50**
Background	36.84	25.93	38.18	38.18	**37.49**	30.88
Cause	00.00	44.44	00.00	13.79	00.00	**21.05**
Comparison	00.00	88.89	00.00	27.59	00.00	**42.11**
Condition	28.21	40.00	73.33	66.67	40.74	**50.00**
Contrast	50.51	47.50	51.55	58.76	51.02	**52.53**
Elaboration	59.63	63.64	83.72	78.49	69.65	**70.29**
Enablement	61.22	63.93	58.82	76.49	59.99	**69.63**
Evaluation	00.00	00.00	00.00	00.00	00.00	00.00
Explanation	40.00	36.36	19.35	17.20	**26.08**	23.35
Joint	51.09	56.36	43.21	57.41	46.82	**56.88**
Manner-Means	00.00	57.14	00.00	66.67	00.00	**61.53**
Summary	00.00	00.00	00.00	00.00	00.00	00.00
Temporal	00.00	28.57	00.00	06.06	00.00	**09.99**
Topic-Comment	00.00	00.00	00.00	00.00	00.00	00.00
Topic-Change	00.00	00.00	00.00	00.00	00.00	00.00
Textual-Organization	00.00	00.00	00.00	00.00	00.00	00.00

TABLE 6 – Scores par relation pour lex expériences sur le RST DT avec la bseline (vote de deux modèles BERT et RoBERTa entraîné sur 10% des données), et le meilleur système obtenu (incluant la baseline et un étiquetage *via* un modèle entraîné sur DiscoGem2. Les étiquettes en gras sont celles présentes dans le corpus DiscoGem2. Les scores en gras sont la meilleure F_1 obtenue.