Alignement de Structures Argumentatives et Discursives par Fouille de Graphes et de Redescriptions

Laurine Huber, Yannick Toussaint, Charlotte Roze, Mathilde Dargnat, Chloé Braud

Alignement de Structures Argumentatives et Discursives par Fouille de Graphes et de Redescriptions.

Laurine Huber*, Yannick Toussaint*
Charlotte Roze*, Mathilde Dargnat**,***, Chloé Braud*

* Université de Lorraine, CNRS, Inria, LORIA (UMR 7503), F-54000 Nancy, France
firstname.lastname@loria.fr
** ATILF, Université de Lorraine, CNRS (UMR 7118), Nancy, France
*** Institut des Sciences Cognitives Marc Jannerod, CNRS (UMR 5304), Bron, France
mathilde.dargnat@univ-lorraine.fr

Résumé. Dans cet article, nous étudions la similarité entre structures argumentatives et discursives en alignant des sous-arbres dans un corpus annoté en RST et en structure argumentative. Contrairement aux travaux précédents, nous ne nous intéressons pas uniquement à un alignement relation à relation, mais à un alignement de sous-structures. À l'aide de méthodes de fouille de données, nous montrons que des similitudes existent entre l'argumentation et le discours. L'annotation multiple du corpus permet également de proposer un alignement entre les structures. De plus, cette approche permet de mettre en évidence les différences d'expressivité des deux formalismes.

1 Introduction

La représentation sémantique d'un texte en Traitement Automatique des Langues se fait à différents niveaux. Le niveau discursif représente sous forme d'un graphe étiqueté les relations sémantico-pragmatiques qui existent entre les segments (i.e clauses ou phrases) d'un texte. Il existe différents cadres théoriques permettant de relier ces segments. L'annotation en RST (Rhetorical Structure Theory, Stede (2008)) représente l'organisation textuelle par le biais de relations de cohérence entre les segments de texte et peut être appliquée à n'importe quel genre textuel. L'annotation de la macro-structure argumentative (notée ARG) proposée par Peldszus et Stede (2013) repose sur les travaux de Freeman (1992). Celle-ci permet de représenter la manière dont les prémisses et les conclusions sont liées au sein d'un texte argumentatif pour former un ensemble cohérent menant à une conclusion principale. Ces deux cadres théoriques ont pour objectif de représenter l'intention de l'auteur par rapport au lecteur mais elles utilisent des ensembles de relations distincts et suivent des règles de construction différentes. Comprendre les liens entre ces deux formalismes permettrait alors de construire des ponts entre les théories et de mieux saisir le pouvoir expressif de chacun d'entre eux. Cet article présente les résultats préliminaires sur l'alignement de ces structures en utilisant la fouille de graphes et la fouille de redescriptions. Nous proposons d'appliquer la fouille de redescriptions sur un corpus de textes annotés en ARG et en RST. Chacune des annotations nous permet de

construire un arbre initial dont nous extrayons les sous-arbres. Ces sous-arbres permettent de construire une vue ARG et une vue RST, e.g. des tables représentant la relation binaire entre l'ensemble des textes et l'ensemble des attributs ARG ou RST respectivement.

Peldszus et Stede (2016) ont montré qu'un alignement deux à deux entre les relations n'est pas toujours possible. Certaines non-correspondances sont expliquées par des différences d'expressivité entre les deux types d'annotation. Nous proposons donc d'étudier un alignement entre sous-structures permettant ainsi la combinaison de relations.

Cabrio et al. (2013) proposent une étude manuelle pour la mise en correspondance de schémas argumentatifs (selon les *Argumentation Schemes* (AS) de Walton et al. (2008)) avec les relations du *Penn Discourse TreeBank* (PDTB) de Prasad et al. (2008) : des correspondances sont conjecturées et 2 annotateurs évaluent leur pertinence. Le coefficient Cohenś Kappa calculé sur les annotations montre un accord significatif qui valide leur hypothèse. Leur approche est basée sur une appréciation et une annotation humaine. Contrairement à eux, nous proposons une approche automatique basée sur la fouille de données. C'est à notre connaissance la première approche automatique et systématique pour l'alignement de structures ARG et RST.

2 Méthodologie

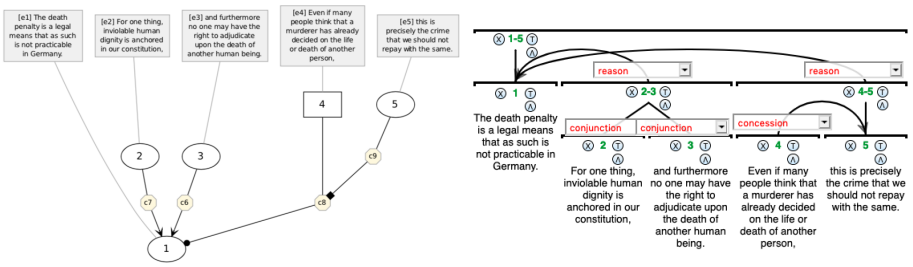
Le processus en trois étapes vise à trouver un alignement exhaustif et systématique dans le corpus entre des "parties" des représentations RST et des "parties" des représentations ARG. Pour toutes les représentations $t \in T$, nous transformons chacune des structures ARG et RST en un arbre A et R (voir Fig. 1) et nous extrayons les sous-arbres $a \in A$ et $r \in R$, produisant ainsi deux ensemble d'attributs distincts. Les deux vues sont ensuite définies par les relations binaires $R_{arg} \subseteq T \times A$ et $R_{rst} \subseteq T \times R$, où $aR_{arg}t$ et $rR_{rst}t$ définissent l'appartenance d'un sous-arbre a à un texte t dans leur représentations ARG et RST respectivement. La fouille de redescriptions permet ensuite d'extraire des paires (q_{Arg}, q_{Rst}) de requêtes, où q_{Arg} est une formule logique construite à partir des attributs de A et q_{Rst} à partir de ceux de R.

Encodage des arbres RST et ARG. Les représentations RST et ARG sont différentes sur plusieurs points (e.g contraintes d'attachement, notion de nucléarité (Stede, 2008)). Cependant, Peldszus et Stede (2016) ont montré que les segments correspondant au noyau principal [1] en RST et la conclusion principale en ARG correspondent dans 85% des textes au même segment. Nous représentons donc pour chaque texte deux arbres initiaux distincts, où la racine (étiquetée `CC`) représente le noyau principal dans la RST et la conclusion dans l'ARG. Pour chaque relation entre la prémisse p et la conclusion c de l'ARG et entre le noyau n et le satellite s de la RST, les relations discursives ou argumentatives correspondent au label de la branche correspondante, et la notion parent-enfant est définie comme suit :

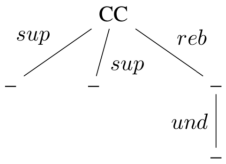
— en ARG, c est le père et p le fils,
— en RST, n est le père et s le fils.

Une particularité existe dans les arbres ARG, où la relation d'`undercut` est dirigée vers un arc et non vers un noeud. Nous nous inspirons de Wachsmuth et al. (2017) et la modifions afin que la cible de la relation `undercut` devienne la prémisse de la relation. La Fig. 1 illustre les annotations ARG et RST et leurs arbres initiaux.

1. L'unité la plus centrale. (Stede, 2008)

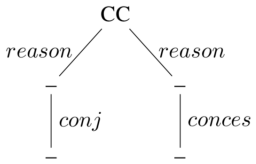


Annotation ARG



Arbre ARG initial

Annotation RST



Arbre RST initial

FIG. 1: Texte *micro_b006* annoté en ARG et en RST et arbres initiaux associés.

Construction des deux vues. Nous extrayons indépendamment les sous-arbres de l'ensemble des arbres RST et ARG. Nous donnons à chacun un identifiant unique, il devient alors un attribut : un texte t possède un attribut $a \in A$ et $r \in R$ si l'arbre correspondant à l'attribut est un sous-arbre de l'arbre initial A de l'ARG ou R de la RST respectivement. Chaque texte possède donc un ensemble d'attributs $a \in A$ et un ensemble d'attributs $r \in R$. Les sous-arbres sont extraits avec gSpan (*Graph-Based Substructure Pattern Mining*) (Yan et Han, 2002), un algorithme qui, étant donné un ensemble de graphes $\mathbb{GS}$, en extrait les sous-graphes fréquents. De façon informelle, un graphe h est un *sous-graphe* de g si h est contenu dans g, et h est *fréquent* si au moins s graphes de $\mathbb{GS}$ contiennent h, s étant un seuil fixé par l'utilisateur [2]. À partir de la sortie de GSpan, nous représentons les attributs booléens dans deux tables binaires, où les lignes correspondent aux textes et les colonnes correspondent aux attributs.

Fouille de redescriptions. En analyse de données, la fouille de redescriptions (Galbrun et Miettinen, 2017) consiste à trouver deux caractérisations différentes d'un même ensemble d'objets (i.e. textes dans notre contexte). L'objectif est de trouver deux expressions $q1$ et $q2$ (des requêtes), où $q1$ et $q2$ sont des formules logiques constituées à partir des attributs $a \in A$ et $r \in R$ respectivement, et où l'ensemble de textes décrits par $q1$ et $q2$ est suffisamment similaire. Cette similarité est mesurée par l'indice de Jaccard J(q1,q2) = $\frac{supp(q1 \wedge q2)}{supp(q1 \vee q2)}$ où $supp(q)$ est le nombre de textes pour lesquels q est vraie. L'indice de Jaccard représente la manière dont se recoupent les objets vrais dans $q1$ et ceux vrais dans $q2$.

2. Nous fixons ce seuil à 2 pour considérer tous les sous-arbres qui sont présents dans au moins deux textes.

La stratégie d'exploration de ReReMi est basée sur la mise à jour atomique. Premièrement, l'algorithme calcule l'indice de Jaccard pour toutes les paires possibles de requêtes atomiques, autrement dit toutes les redescriptions qui peuvent être construites à partir d'un attribut de chaque vue. Les n meilleures paires sont conservées. A partir de ces paires, l'algorithme applique des opérations d'addition sur l'une et l'autre des requêtes afin d'améliorer les redescriptions candidates jusqu'à ce que l'indice de Jaccard ne puisse plus être amélioré. Nous utilisons l'algorithme ReReMi (Galbrun et Miettinen, 2012) implémenté dans l'outil Siren (Galbrun et Miettinen, 2018) avec les paramètres prédéfinis par l'outil. Les conjonctions et les disjonctions sont autorisées dans les requêtes mais la longueur des requêtes est limitée à quatre attributs.

3 Expérimentation

Données. Le corpus est composé de 112 textes répondant à une question controversée (e.g *"Should Germany introduce the death penalty ?"*). Les arbres RST sont annotés avec 28 relations distinctes, chacun des textes ayant entre 2 et 12 relations par arbre. Les relations RST les plus fréquentes sont : `reason` (180), `concession` (65), `list` (63), `conjunction` (44), `antithesis` (32), `elaboration` (37), et `cause` (20). Cinq relations distinctes servent à annoter les arbres ARG, et chaque texte possède entre 2 et 9 relations. Les relations ARG les plus fréquentes sont : `support` (263), `rebut` (108) et `undercut` (63). Avec gSpan, nous avons extrait 311 sous-arbres RST et 98 sous-arbres ARG. L'attribut RST le plus fréquent apparaît dans 105 textes alors que l'attribut ARG le plus fréquent apparaît dans 94 textes. Seuls 22 attributs RST sont partagés par plus de 10 textes, et 18 attributs ARG sont partagés par plus de 13 textes.

	q1	q2	J(q1,q2)	# texts
$Rd1$	`a57`	`r40` ∨ `r65` ∨ `r123`	0.691	54
$Rd2$	`a58`	`r61` ∨ `r119` ∨ `r125`	0.351	13
$Rd3$	`a23` ∨ `a59`	`r125`	0.3	8

TAB. 1: 3 des 31 redescriptions. `aX` et `rX` correspondent resp. aux sous-arbres ARG et RST.

Résultats. Pour des raisons de place, nous ne commenterons que 3 des 35 redescriptions obtenues (voir Tab. 1). Nous choisissons $Rd1$ car elle a l'indice de Jaccard le plus haut, $Rd2$ car c'est une spécialisation de $Rd1$ et $Rd3$ car celle-ci contient une disjonction du côté ARG. Les attributs de chacune des redescriptions sont représentés en Fig. 2, Fig. 3 et Fig. 4.

Les 54 textes décrits par $Rd1$ contiennent tous l'attribut `a57` en ARG, mais la disjonction côté RST met au jour une différence de granularité entre les deux formalismes. Plus précisément, parmi les 54 arbres qui ont `a57`, 30 contiennent `r123`, 22 contiennent `r65`, 2 contiennent `r40` dans leur représentation RST. En d'autres termes, près de la moitié du corpus contient deux relations de `support` dirigées vers la CC, et ces textes ont dans leur arbre ARG soit une relation `reason` avec 2 éléments en `list`, soit deux relations `reason`, soit une `motivation` dirigée vers la `CC`. Les objets décrits par $Rd2$ et $Rd3$ sont aussi décrits par $Rd1$ donc $Rd2$ et $Rd3$ peuvent être vus comme des spécialisations de $Rd1$. $Rd2$ peut être lue de la même manière que $Rd1$: parmi les 23 textes qui contiennent `a58`, 13 sont alignés avec soit `r61` (3), soit `r119` (3), soit `r125` (7).

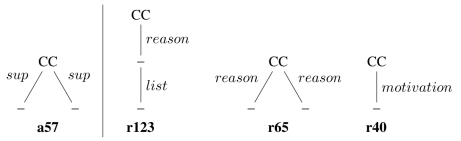


FIG. 2: Sous-arbres correspondants aux attributs de $Rd1$

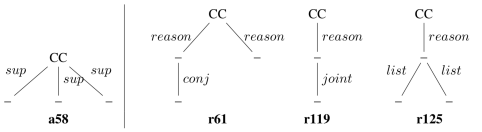


FIG. 3: Sous-arbres correspondants aux attributs de $Rd2$

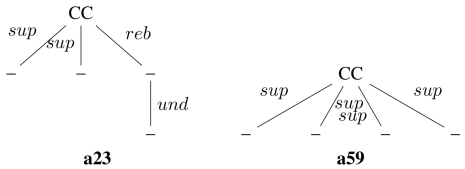


FIG. 4: Sous-arbres correspondants aux attributs de $Rd3$

Ces deux premières redescriptions confirment qu'une relation de `support` en ARG peut correspondre à différentes relations en RST (Peldszus et Stede (2016)). Notre approche permet cependant de révéler les attributs RST qui apparaissent en vis-à-vis d'une structure ARG et d'en quantifier le nombre d'occurences. Contrairement à $Rd1$ et $Rd2$, la disjonction du côté de l'ARG dans $Rd3$ suggère que l'attribut `r125` (qui apparaît dans 8 textes) peut être aligné avec deux différentes structures ARG : `a59` (dans 2 textes), et `a23` (dans 5 textes). Bien qu'ayant un faible indice de Jaccard, cette redescription est pertinente si il existe une relation `undercut` entre X et Y, et une `rebut` entre Y et la CC, alors X est en fait un argument en `support` de la CC. Néanmoins, l'approche engendre aussi des alignements incohérents dus à l'anonymisation des segments de texte et aux paramètres de ReRemi.

4 Conclusion

La fouille de redescriptions peut être appliquée sur des sous-arbres pour aligner différentes structures textuelles. Ce processus automatique vise à proposer une comparaison systématique de différents formalismes. Appliquée à un corpus annoté en ARG et RST, cette expérience préliminaire permet de mettre en exergue des différences de granularité et d'encodage entre les formalismes. Les prochains travaux doivent permettre de préserver l'ancrage sur les segments textuels et d'étendre les expériences à d'autres formalismes (par exemple la SDRT).

Références

Cabrio, E., S. Tonelli, et S. Villata (2013). From Discourse Analysis to Argumentation Schemes and Back : Relations and Differences. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Leite, T. C. Son, P. Torroni,

L. van der Torre, et S. Woltran (Eds.), *Computational Logic in Multi-Agent Systems*, Volume 8143, pp. 1–17. Berlin, Heidelberg : Springer.

Freeman, J. B. (1992). *Dialectics and the Macrostructure of Argument*. Berlin : Foris.

Galbrun, E. et P. Miettinen (2012). From black and white to full color : extending redescription mining outside the Boolean world. *Statistical Analysis and Data Mining : The ASA Data Science Journal 5*(4), 284–303.

Galbrun, E. et P. Miettinen (2017). *Redescription Mining*. Springer International Publishing.

Galbrun, E. et P. Miettinen (2018). Mining redescriptions with Siren. *ACM Transactions on Knowledge Discovery from Data (TKDD) 12*(1), 6 :1–6 :30.

Peldszus, A. et M. Stede (2013). From Argument Diagrams to Argumentation Mining in Texts : A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 7*(1), 1–31.

Peldszus, A. et M. Stede (2016). Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, Berlin, Germany, pp. 103–112. Association for Computational Linguistics.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, et B. Webber (2008). The Penn discourse treebank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Stede, M. (2008). Rst revisited : Disentangling nuclearity. *'Subordination' versus 'Coordination' in Sentence and Text*, 33–59.

Wachsmuth, H., G. Da San Martino, D. Kiesel, et B. Stein (2017). The impact of modeling overall argumentation with tree kernels. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2379–2389. Association for Computational Linguistics.

Walton, D., C. Reed, et F. Macagno (2008). *Argumentation Schemes*. Cambridge : Cambridge University Press.

Yan, X. et J. Han (2002). gSpan : graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, Maebashi City, Japan, pp. 721–724. IEEE.

Summary

In this paper, we investigate similarities between discourse and argumentation structures by aligning subtrees in a corpus containing both annotations. Contrary to previous works, we focus on comparing sub-structures and not only relation matches. Using data mining techniques, we show that discourse and argumentation most often align well, and the double annotation allows to derive a mapping between structures. Moreover, this approach enables the study of similarities between discourse structures and differences in their expressive power.