Apprentissage faiblement supervisé de la structure discursive

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, Nicholas Asher

Apprentissage faiblement supervisé de la structure discursive

Sonia Badene [1,2] Kate Thompson [1] Jean-Pierre Lorré [2] Nicholas Asher [1]
(1) IRIT, Université de Toulouse, (2) LINAGORA
soniabadene@gmail.com, kate.thompson@irit.fr, jplorre@linagora.com,
nicholas.asher@irit.fr

RÉSUMÉ

L'avènement des techniques d'apprentissage automatique profond a fait naître un besoin énorme de données d'entraînement. De telles données d'entraînement sont extrêmement coûteuses à créer, surtout lorsqu'une expertise dans le domaine est requise. L'une de ces tâches est l'apprentissage de la structure sémantique du discours, tâche très complexe avec des structures récursives avec des données éparses, mais qui est essentielle pour extraire des informations sémantiques profondes du texte. Nous décrivons nos expérimentations sur l'attachement des unités discursives pour former une structure, en utilisant le paradigme du *data programming* dans lequel peu ou pas d'annotations sont utilisées pour construire un ensemble de données d'entraînement "bruité". Le corpus de dialogues utilisé illustre des contraintes à la fois linguistiques et non-linguistiques intéressantes qui doivent être apprises. Nous nous concentrons sur la structure des règles utilisées pour construire un modèle génératif et montrons la compétitivité de notre approche par rapport à l'apprentissage supervisé classique.

ABSTRACT

Learning discourse structure using weak supervision

The advent of Deep Learning techniques has created a critical need for more labeled training data. Such training data are extremely expensive to create, especially when domain expertise is required. One such task is the learning of semantic structure and discourse structure, which are typically very complex involving recursion but which are essential for extracting deep semantic information from text. In this article, we will show our experiments with the *data programming* paradigm, in which few to no annotations are used to build the data set for the attachment problem in discourse, the first step in forming the complete structure of discourse. The corpus of situated dialogues we use exhibits interesting structural constraints on both linguistic and non-linguistic components that need to be learned. We focus on the rules structure used to build the generative model and show the competitiveness of our approach compared to traditional supervised learning.

MOTS-CLÉS : Structure du discours, Supervision distante, Attachement, *Data Programming* .

KEYWORDS: Discourse Structure, Weak Supervision, Attachment, Data Programming .

1 Introduction

L'arrivée des nouvelles méthodes d'apprentissage profond a beaucoup simplifié l'étape d'extraction de caractéristiques de données. Cependant, pour que ces techniques basées sur des algorithmes d'apprentissage supervisé puissent apprendre ces attributs, il faut avoir beaucoup de données étiquetées, des données d'apprentissage sur lesquelles ils peuvent être formés. L'étiquetage manuel des données est à la fois coûteux et long, surtout lorsqu'une expertise dans le domaine est requise ou qu'ultérieurement il faut ré-étiqueter d'une nouvelle manière les données.

Dans cet article, nous montrerons comment Snorkel (Ratner *et al.*, 2017), un nouveau paradigme de programmation par les données (*data programming* en anglais) peut construire un ensemble suffisant de données d'entraînement pour résoudre l'attachement, un problème clef dans le processus d'inférence d'une structure pour un discours. La méthode de Snorkel ne perd que 4% d'exactitude par rapport à une méthode classique d'entraînement avec des données manuellement annotées.

Plus important encore, nous montrons que dans le cadre de Snorkel on peut construire un "modèle génératif" qui est en effet *plus performant* que notre modèle classique. Les probabilités du modèle génératif peuvent aussi servir d'entrée à un modèle discriminatif. Snorkel apporte un cadre pour fournir des informations symboliques de manière efficace à un processus connexionniste ou statistique qui doit généraliser et lisser les résultats fournis par cette partie symbolique, exemplifiant ainsi une IA "hybride" employant des représentations symboliques et des méthodes connexionnistes.

2 État de l'art

Il existe plusieurs théories de la structure du discours pour les textes : RST (Rhetorical Structure Theory) (Mann & Thompson, 1987), LDM (Linguistic Discourse Model) (Polanyi *et al.*, 2004), PDTB (The Penn Discourse Treebank) (Prasad *et al.*, 2007) et SDRT (Segmented Discourse Representation Theory) (Asher & Lascarides, 2003). Bien que le travail d'analyse du discours soit largement concentré sur la théorie de la structure rhétorique (RST), comme démontré par Morey *et al.* (2018), les structures discursives ont intérêt à être traduites dans des arbres de dépendance (Muller *et al.*, 2012; Afantenos *et al.*, 2015). De plus, nous nous intéressons à l'étude de dialogues multi-locuteurs pour lesquels le seul corpus annoté est STAC [1] (Asher *et al.*, 2016). Dans ce corpus on trouve une portion significative de structures qui sont naturellement interprétées de façon non arborescente, ce qui exclurait un traitement en termes de DLTAG, LDM ou RST. Nous partons sur la base de cette discussion et travaillons avec des structures de la SDRT simplifiées et nous nous baserons sur les initiatives de Perret *et al.* (2016) où les structures discursives attendues sont des graphes. [2]

Dans cet article, nous proposons la création des données d'entraînement pour la tâche de prédiction d'attachement dans le corpus de dialogues STAC (Asher *et al.*, 2019). Pour cela, nous utilisons la *programmation par les données*, un paradigme pour la création et la modélisation des données d'entraînement. La programmation par les données fournit un cadre simple et unificateur pour une faible supervision, dans lequel les étiquettes d'entraînement sont bruitées et peuvent provenir de sources multiples et potentiellement contradictoires. Dans ce cadre, on peut coder cette faible supervision sous la forme de fonctions d'étiquetage, qui fournissent chacune une étiquette pour un sous-ensemble de données. De nombreuses approches de supervision faibles différentes peuvent être exprimées sous forme de fonctions d'étiquetage, telles que les stratégies qui utilisent des bases de connaissances existantes, comme dans la supervision distante (Mintz *et al.*, 2009).

3 Expérimentations

3.1 Corpus de dialogues annotés

Pour notre expérimentation nous avons utilisé STAC, un corpus de discussions spontanées autour du jeu *Colons de Catan* de négociations multi-locuteurs annotées pour la structure discursive dans le style

1. Lien vers le corpus STAC : https ://www.irit.fr/STAC/index.html

2. Les structures de la SDRT ont une complexité difficile à gérer–les structures complexes ou CDUs (Complexe Discourse Units) (Venant *et al.*, 2013). Comme d'autres travaux sur les structures discursives en SDRT (Muller *et al.*, 2012; Perret *et al.*, 2016), nous simplifions cette structure des CDUs (voir Section 3.2).

de la SDRT. Cette annotation inclut des coups linguistiques mais aussi des actions non-linguistiques comme l'action de terminer son tour de négociation ou de construire une route comme illustrée sur la figure 1. Nous avons choisi ce corpus qui est annoté manuellement sur l'attachement afin d'évaluer notre approche, mais aussi parce que l'analyse des corpus de dialogues est de plus en plus demandée avec l'arrivée des assistants conversationnels, des chatbots ou des corpus de plateforme de discussion en ligne.

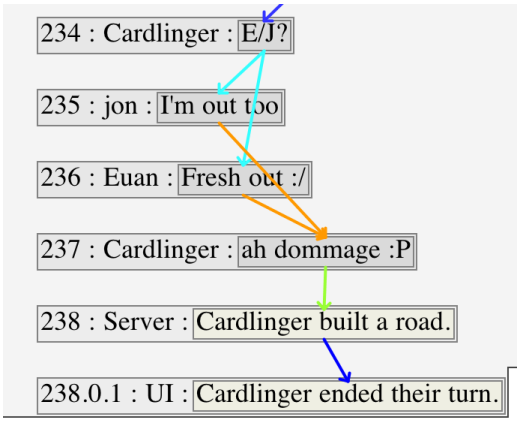

FIGURE 1 – Extrait d'un dialogue sur STAC qui montre des relations comme Sequence (bleu foncé), Result (vert), et les coups linguistiques (énoncés par les joueurs) et non-linguistiques (donnés par le "Serveur" ou "UI")

Le corpus présente des contraintes structurelles intéressantes sur les composantes linguistiques et non linguistiques qui doivent être apprises. Le corpus non-linguistique est très régulier, donc facile à modéliser. Par contre, la structure discursive entre tours linguistiques, ou entre tours non-linguistique et tours linguistiques est riche et difficile à capturer. Dans notre étude, on s'intéresse aux tours linguistiques et aux tours qui marquent une transition entre le linguistique et les coups non-linguistiques.

3.2 Cadre expérimental

Avant de commencer nos expériences, nous avons mis en œuvre les prétraitements suivants :

1. La complexité des structures annotées de la SDRT étant difficile à prédire, nous avons suivis les travaux de (Muller *et al.*, 2012; Perret *et al.*, 2016) et avons mis en place un algorithme simple de "flattening" ("aplatissement") afin de remplacer les CDUs par des relations entre paires d'unités élémentaires discursives. Un CDU est un graphe de dépendance avec plusieurs DU/segments comme sommets. Cet ensemble est considéré comme un segment qui peut être reliée à d'autres segments. Pour aplatir les graphes, pour chaque CDU nous avons identifié la "tête" du CDU, qui est le premier segment du CDU, et connectons toutes les relations entrantes et sortantes du CDU à la tête - pour nos 4 types de relations, il y a un total de 33 865 relations dans le corpus (incluant celles des dialogues non linguistiques), et environ 40% de celles-ci ont été ajustées, soit au noeud source, soit au noeud cible (ou aux deux).
2. Nous avons également restreint les dialogues dans le corpus afin d'étudier essentiellement les dialogues linguistiques. Nous avons éliminer tous les dialogues qui n'ont pas de conversations linguistiques - 1463 dialogues ont été supprimés, ce qui nous laisse 1130 dialogues qui contiennent 18 767 segments non linguistiques et 13 734 segments linguistiques, avec 31 251 relations.

3. Nous avons aussi ignoré tous les attachements qui ont une distance supérieure à 10 (c'est-à-dire qui ont plus de 9 unités élémentaires discursives entre la source et la cible de l'attachement). Les distances relationnelles varient de 1 à 160 dans le corpus de développement. 67% des relations ayant la distance 1, et 98% des relations ayant la distance 10 ou moins.
4. Pour cette première tâche de prédiction d'attachement, nous avons travaillé avec 4 types de relations les plus fréquentes : Question-answer-pair (QAP), Sequence (temporelle), Result (relation causale), Continuation (la relation de continuité thématique) - ce qui représente 70% des relations dans le corpus.
5. Afin de réduire le temps d'exécution de chaque règle pendant le développement, nous avons créé des ensembles restreints pour chaque type de relation : des versions plus petites de l'ensemble de développement qui ignorent toutes les paires qui ne pouvaient pas être attachées par le type de relation en question. Nous disposons de sous-ensembles de données pour chaque relation discursive particulière et d'un ensemble plus vaste des données pour les règles des quatre relations discursives que nous avons examinées.

3.3 Les différentes étapes du traitement

3.3.1 Les candidats et les Fonctions d'étiquetage

Les candidats sont les entrées pour lesquelles les étiquettes seront prédites et ils sont extraits des données en fonction de la nature de la tâche de prédiction. Puisque nous prédisons l'attachement entre deux segments d'un dialogue, nos candidats sont l'ensemble de toutes les combinaisons possibles *source-cible* des segments dans chaque dialogue, limitées par une distance relationnelle maximale de 10. Le corpus STAC nous donne déjà la segmentation en unités discursives élémentaires des dialogues du corpus. Nous avons construit notre propre algorithme pour générer les candidats en entrée pour notre modèle génératif. Pour chaque dialogue, nous avons créé une liste de toutes les paires uniques possibles. Comme certaines relations discursives présentent des liens vers l'arrière (les 4 relations étudiées dans cet article n'ayant pas de relations vers l'arrière, nous avons tout de même généré ces candidats afin d'évaluer l'attachement sur la structure globale), nous avons ajouté des contraintes comme dans (Perret *et al.*, 2016) afin d'éviter l'explosion combinatoire de candidats.

Dans Snorkel, les règles s'appellent des fonctions de labellisation (FL). Les FL sont appliquées à chaque candidat et retourne un "1", "-1" ou "0" qui signifie que les deux segments du candidat sont "liés" , "non-liés", ou "on ne sait pas". Ces fonctions utilisent les informations "locales" des candidats (le texte avec la syntaxe, les connecteurs ...), y compris les identités des interlocuteurs et des destinataires, les actes de dialogue, les types des segments (linguistique ou non-linguistique) et la distance entre les segments, afin de saisir le modèle général sous-tendant. Comme nous l'avons vu plus haut, nous prédisons l'attachement en aillant en tête les 4 types de relations – *Result*, *QAP*, *Continuation* et *Sequence*. De cette façon, les FL utilisent également une information de type. Cela a du sens d'un point de vue à la fois empirique et épistémologique : une décision d'attachement discursive entre deux segments est étroitement liée au type de la relation qui les lie, et donc lorsqu'un annotateur décide que deux éléments discursifs élémentaires sont attachés, il ou elle le fait avec une certaine connaissance du type de relation qui les lie. La figure 2 montre un exemple de nos règles utilisées pour la prédiction d'attachement avec la relation *Result* en tête.

Si nous nous concentrons sur l'information locale en construisant les FL, le besoin de s'appuyer sur l'information considérée globale devient évident du fait que, si nous n'avons pas de moyen de surveiller et noter où nous sommes exactement dans un dialogue et où sont les attachements déjà

```python
def LF_Result_L_L_case1(row):
    l=0
    if (any(x in row.target_text.lower() for x in resultWords)
                or any(x in row.source_text.lower() for x in resultWords)):
            l=1
    return l


def LF_Result_L_L_case2(row):
    l = 0
    if row.source_surface_act in ["Question", "Request", "Assertion"] \
    and (row.target_dialogue_act in ["Offer", "Counteroffer"] \
         or row.source_emitter == row.target_text.partition(' ')[0] or row.target_surface_act == "Request"):
        l=1
    return l
```

FIGURE 2 – *Result* relie une cause à son effet. Voici un exemple de nos règles écrites en python pour la relation *Result* reliant deux unités de discours linguistiques.

prédits, nous risquons de sur-étiqueter les candidats “liés”. Ce qui est très inefficace dans un corpus dont les données sont éparses. Ainsi, nous avons ordonné les candidats pour appliquer les FL aux candidats des segments adjacents en premier, et puis regarder ceux des segments de plus en plus éloignés, tout en maintenant une liste de tous les segments déjà prédits “lié”. Ces mesures prises nous permettent de nous servir des faits contextuels simples, et donc de construire des FL plus sensibles au contexte, ce qui aboutit à une différence de 5 points d’exactitude de nos règles par rapport aux exemples dans le corpus de développement.

3.3.2 Le modèle génératif

Une fois que nous appliquons l’ensemble des FL à tous les candidats, nous passons à l’étape "générative". Dans le système Snorkel, le modèle génératif unit les résultats des FL : une matrice des étiquettes donnée par chaque FL (colonnes) sur les candidats (lignes) est alors générée. Bien que l’approche la plus simple serait de prendre le vote majoritaire entre les FL pour chaque candidat, celle-ci serait moins efficace dans les situations où nous n’avons pas beaucoup de votes sur une entrée, ou si toutes les FL s’abstiennent. De plus, elle ne prendrait pas en compte les performances individuelles des FL. Donc pour apporter des améliorations au vote majoritaire, le modèle génératif cherche à maximiser la probabilité marginale des FL de chaque candidat pour apprendre une estimation des précisions des FL et les pondérer selon ces précisions (Bach *et al.*, 2017). Ensuite, le modèle calcule pour chaque candidat la probabilité d’être “1” ou “0” (“lié” ou “non-lié”) dans le contexte de notre tâche de prédiction binaire.

Ce calcul suppose que les FL sont indépendantes. Cependant, les FL fournies sont souvent dépendantes : par exemple, les FL peuvent être de simples variations les unes des autres ou peuvent dépendre d’une source commune de supervision distante. Si nous ne tenons pas compte des dépendances entre les fonctions d’étiquetage, nous pouvons avoir toutes sortes de problèmes. La méthode de sélection automatique des dépendances à modéliser dans Snorkel, sans accès aux données de référence, utilise un estimateur de pseudo-vraisemblance, qui ne nécessite aucun échantillonnage ni approximation pour calculer le gradient objectif et ceci est plus rapide que l’estimation du maximum de vraisemblance. Cela évite d’indiquer les dépendances à la main, tâche difficile et sujet aux erreurs.

3.3.3 Un modèle discriminatif de référence

Alors que le modèle génératif est essentiellement une combinaison pondérée des FL fournies par l'utilisateur - qui ont tendance à être précises mais à faible couverture-, le modèle discriminatif peut conserver cette précision tout en apprenant à généraliser au-delà des fonctions d'étiquetage, augmentant ainsi la couverture et la robustesse sur les données non encore visualisées. Le rappel est alors plus élevé dans la plupart des cas, même si parfois on observe une petite baisse de précision.

Nous avons utilisé le modèle de classification séquentielle de BERT (Devlin *et al.*, 2018) (code source sur le lien ci-dessous [3]) avec 10 époques pour l'entraînement et tous les paramètres par défaut. BERT, Bidirectional Encoder Representations from Transformers, est un "encoder" de texte entraîné à l'aide de modèles de langage où le système doit deviner un mot manquant ou un élément de mot qui est supprimé au hasard du texte. Conçue à l'origine pour les tâches de traduction automatique, BERT utilise l'auto-attention bidirectionnelle pour produire les encodages et produit des résultats qui dépassent l'état de l'art sur de nombreuses tâches de classification textuelle. Alors qu'en principe, nous aurions pu utiliser n'importe quel modèle discriminatif, comme le suggère la littérature de Snorkel, BERT nous a donné de loin les meilleurs résultats sur la prédiction de l'attachement. C'est pourquoi nous avons également utilisé BERT comme modèle pour l'apprentissage supervisé de l'attachement afin de comparer ses résultats avec ceux de la méthode de supervision faible.

4 Résultats et analyse

	VP	VN	FP	FN	Exactitude
QAP LL	294	1798	112	138	0.89
QAP NLNL	84	187	0	0	1
RES NLNL	739	2929	13	55	0.98
RES LNL	13	2158	93	97	0.91
RES LL	25	316	19	37	0.85
RES NLL	2	139	0	2	0.98
Cont LL	16	9818	110	106	0.97
Cont NLNL	613	3254	0	1	0.99
SEQ NLL	90	658	2	14	0.97
SEQ NLNL	236	1220	10	76	0.94

TABLE 1 – Nombre de vrais positifs (VP), de vrais négatifs(VN), de faux positifs (FP) et de faux négatifs (FN) pour chacune de nos fonctions d'étiquetage lorsqu'elles sont appliqués aux candidats associés aux types des relations discursives utilisées.

Nous avons d'abord évalué les FL pour chaque type de relation discursive individuellement sur les sous-corpus de développement, en fournissant une mesure de leur couverture et de leur précision (Tableau 1). Ensuite, nous avons évalué le modèle génératif sur la combinaison des quatre types de FL. Le tableau 2 présente les résultats à la fin de chaque étape de notre système de supervision faible (Modèle Génératif). Pour comparer les deux approches, le modèle discriminatif a été entraîné sur les marginaux fournis par notre modèle génératif, mais aussi sur les annotations manuelles du corpus Stac. L'évaluation du modèle discriminatif s'est faite sur l'ensemble test du corpus.

Les résultats du modèle génératif sur l'attachement sont près de 20 points plus élevés en F1 mesure

3. Lien vers le code source du modèle de classification séquentielle de BERT : https ://github.com/huggingface/pytorch-pretrained-BERT/blob/master/examples/run_classifier.py

	Modèle Génératif			Modèle Discriminatif sur Test	
	Dev	Train	Test	avec Marginales	avec annotations Manuelles
Précision	0.67	0.70	0.68	0.45	0.61
Rappel	0.84	0.85	0.84	0.54	0.53
F1 mesure	0.75	0.77	0.75	0.49	0.57
Exactitude	0.92	0.93	0.92	0.84	0.88

TABLE 2 – Évaluations de prédiction de l'attachement avec la combinaison de toutes les règles des quatre types modélisées dans cet article, avec les approches faiblement supervisées et supervisées.

par rapport au modèle discriminatif entraîné sur les annotations manuelles. Cela montre la puissance de l'approche fondée sur des règles et la supervision faible, même lorsqu'on la compare à un système d'apprentissage profond à l'état de l'art. Un autre point intéressant est que le modèle discriminatif a des résultats acceptables avec les données marginales par rapport à sa performance en utilisant les annotations manuelles ; son exactitude n'est inférieure que de 4 points et son score F1 est inférieur de 8 points mais toujours comparable aux résultats de la littérature, montrant que le modèle génératif transmet bien des informations au modèle discriminatif. Plutôt que de traiter naïvement ces étiquettes bruyantes comme une vérité de base, notre modèle discriminatif sensible au bruit donne une légère amélioration dans le rappel avec une diminution de la précision par rapport à l'approche supervisée. En ce qui concerne les FL individuelles isolées, nous constatons qu'à part *QAP*, nos règles pour chaque type de relation ont une exactitude, une précision et un rappel comparables à ceux des modèles supervisés. L'une des raisons de notre précision plus faible pour *QAP* peut être attribuée aux conséquences de la procédure d'aplatissement réalisée en pré-traitement ; dans certains cas, l'algorithme d'aplatissement rattache la relation *QAP* à la tête d'un CDU qui en fait n'était pas le segment du CDU qui a marqué la question. Ce qui est intéressant, c'est la synergie entre les règles, de sorte que lorsqu'elles interagissent toutes sur les données de test, elles réussissent très bien sur le modèle génératif.

5 Conclusion et perspectives

Ayant choisi un modèle discriminatif unique pour toutes nos expérimentations, nous avons pu comparer notre approche hybride utilisant Snorkel avec celle du modèle classique sur une tâche difficile, celle de l'attachement discursive. Notre approche permet de modéliser plus finement le discours et d'être généralisée à d'autres corpus. Contrairement à un algorithme supervisé, nos résultats sur le modèle génératif sont supérieurs de près de 30 points sans couvrir tous les types de règles. Nous générons ainsi beaucoup de données annotées en très peu de temps. Comme perspectives nous envisageons d'enrichir notre modèle Snorkel d'abord en couvrants tous les types de relations et en implémentant des règles qui prendront en compte les contraintes de structuration globale, et non seulement au niveau des paires d'éléments discursives élémentaires comme réalisé jusqu'à présent.

Remerciements

Ce travail a été réalisé dans le cadre du projet de recherche PIA Grands Défis du Numérique LinTO-Assistant vocal open-source respectueux des données personnelles pour l'entreprise- soutenu par Bpifrance N°P169201.

Références

AFANTENOS S., KOW E., ASHER N. & PERRET J. (2015). : Association for Computational Linguistics (ACL).

ASHER N., HUNTER J., MOREY M., BENAMARA F. & AFANTENOS S. D. (2016). Discourse structure and dialogue acts in multiparty dialogue : the stac corpus. In *LREC*.

ASHER N., HUNTER J. & THOMPSON C. (2019). Comparing discourse structures between purely linguistic and situated messages in an annotated corpus. submitted.

ASHER N. & LASCARIDES A. (2003). *Logics of conversation*. Cambridge University Press.

BACH S. H., HE B., RATNER A. & RÉ C. (2017). Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, p. 273–282 : JMLR. org.

F. BENAMARA, N. HATHOUT, P. MULLER & S. OZDOWSKA, Eds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

G. DIAS, Ed. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

MANN W. C. & THOMPSON S. A. (1987). Rhetorical structure theory : Description and construction of text structures. In *Natural language generation*, p. 85–95. Springer.

MINTZ M., BILLS S., SNOW R. & JURAFSKY D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, p. 1003–1011 : Association for Computational Linguistics.

MOREY M., MULLER P. & ASHER N. (2018). A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics*, p. 198–235.

MULLER P., AFANTENOS S., DENIS P. & ASHER N. (2012). Constrained decoding for text-level discourse parsing. *Proceedings of COLING 2012*, p. 1883–1900.

PERRET J., AFANTENOS S., ASHER N. & MOREY M. (2016). Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 99–109.

POLANYI L., CULY C., VAN DEN BERG M., THIONE G. L. & AHN D. (2004). A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

PRASAD R., MILTSAKAKI E., DINESH N., LEE A., JOSHI A., ROBALDO L. & WEBBER B. (2007). The penn discourse treebank 2.0. annotation manual. the pdtb research group.

RATNER A., BACH S. H., EHRENBERG H., FRIES J., WU S. & RÉ C. (2017). Snorkel : Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, **11**(3), 269–282.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

VENANT A., ASHER N., MULLER P., DENIS P. & AFANTENOS S. (2013). Expressivity and comparison of models of discourse structure. In *Proceedings of the SIGDIAL 2013 Conference*, p. 2–11.