Synthèse bibliographique

Traoré Kalifou et Emmanuel de Bézenac
en M1 DAC module TAL

INTRODUCTION

L'analyse du discours dans le contexte du traitement automatique du langage a pour but une compréhension plus fine et plus en profondeur du texte en tentant de dégager une structure discursive dans celui-ci. Certaines subtilités sont difficilement détectables avec la plupart des modèles utilisés en traitement automatique du langage, qui se basent essentiellement sur des traits surfaçiques comme le nombre d'occurrences d'un mot dans un document par exemple. Prenons un exemple qui illustre l'importance de pouvoir exploiter la structure discursive d'un texte trouvé dans *[NLPERS]*:

1: J'aime uniquement voyager en Europe. Alors j'ai soumis un article à ACL.
2. J'aime uniquement voyager en Europe. Donc j'ai soumis une article à ACL.

Lorsque ces deux couples de phrases sont traités indépendemment, on peut inférer les mêmes connaissances: *Il aime voyager en Europe*, *et il a soumis un article à ACL*. Mais grâce à l'analyse du discours nous pouvons inférer davantage d'informations, à savoir, en 1: *ACL se déroule en Europe*, et en 2: *ACL n'est pas en Europe.*

Plus concrètement, l'analyse discursive vise d'abord à segmenter le texte en unités élémentaire du discours (EDU), puis à les rattacher, et identifier leur type de relation, puis récursivement les paires attachées sont liées à des segments simples ou complexes pour aboutir à une structure couvrant le document. Voici un exemple donné par *[Braud Denis]*:

Exemple 1.1: {{[La hulotte est un rapace nocturne] [mais elle peut vivre le jour.]} contrast [La hulotte mesure une quarantaine de centimètres.]} continuation

Exemple 1.2: {[Juliette est tombée.] [Marion l'a poussée.]} explanation

Il existe différentes méthodes pour essayer de répondre à ce problème. C'est un problème reconnu comme difficile; la distinction entre les relations du discours nécessitent un jugement sémantique subtil difficilement capturable en utilisant des features standards. La plupart des travaux étudiés se basent sur des théories du discours, telles que la Rhetorical Structure Theory (RST) *[Mann and Thompson, 1988]* et la Segmented Discourse Representation Theory (SDRT) *[Asher et Lascarides, 2003]* pour définir l'analyse du discours dans le contexte du TAL. Il y a néanmoins peu de travaux sur le sujet avant le début des années 2000. *[Daniel Marcu, Magdalena Romera, and Estibaliz Amorrortu (1999)]* relatant les difficultés qu'ils rencontrent lors de la création d'un corpus d'arbres discursifs afin de constituer une base d'apprentissage. En 2002, avec apparaît le RST-TreeBank, un corpus d'apprentissage encore utilisé dans l'état de l'art *[Lynn*

Carlson, Daniel Marcu, and Mary Ellen Okurowski (2002)]. Aujourd'hui la construction automatique d'une telle structure de dépendances entre EDU utilise essentiellement des modèles d'apprentissage automatique.

Voici une synthèse de certains procédés utilisés par des chercheurs dans le domaine. Nous nous sommes données comme base de travail quatre articles, et puis nous sommes allés chercher des informations dans d'autres articles ou de d'autres sources. Voici les quatre articles que nous avons approfondis, ainsi que leurs auteurs:

- [Braud Denis 2013] Identification automatique des relations discursives «implicites» à partir de données annotées et de corpus bruts, 2013, **Chloé Braud**, ALPAGE, INRIA Paris-Rocquencourt & Université Paris Diderot, **Pascal Denis**, MAGNET, INRIA Lille Nord-Europe.
- [Ji Eisenstein 2014] Representation Learning for Text-level Discourse Parsing, 2014, **Yangfeng Ji**, School of Interactive Computing , Georgia Institute of Technology INRIA, and **Jacob Eisenstein**, School of Interactive Computing, Georgia Institute of Technology.
- [Li Wang Cao Li 2014] Text-level Discourse Dependency Parsing, **Sujian Li**, **Liang Wang**, **Ziqiang Cao**, et **Wenjie Li**,2014
- [Li Rumeng Li Hovy 2014] Recursive Deep Models for Discourse Parsing, **Jiwei Li**, **Rumeng**, **Li** et **Eduard Hovy**, 2014

Nous allons d'abord parler des différentes manières d'aborder le problème de l'analyse du discours, puis des données utilisés et des problèmes associés, des modèles d'apprentissage utilisés, et pour finir, les résultats expérimentaux des différents modèles, ainsi qu'une analyse de ces derniers.

I - CORPUS

Se baser sur un corpus étiqueté est nécessaire vu que les modèles d'apprentissage sont supervisés. Que ce soit pour le Français ou l'Anglais, il existe des corpus de discours annotés. En Anglais, le corpus le plus couramment utilisé est le RST-Treebank, ou encore le Penn discourse Treebank *[Prasad et al., 2008]* . Ce dernier est un corpus constitué d'un million d'articles annotés du Wall Street Journal segmentés en EDU. Les relations entre EDU sont annotées à l'aide de relations discursives qui sont catégorisées et regroupées sous formes de classes (exemple: *TEMPORAL*, qui contient *Asynchronous, Synchronous*, qui lui-même contient *precedence,* et *succession*). Le RST-Treebank est le corpus utilisé par la quasi-totalité des articles que nous avons étudiés. Il est composé d'un corpus de 385 articles de Wall Street Journal du Penn Discourse Treebank. Il est constitué de 176 000 mots, contenant chacun entre 458 et 21 789 mots. Pour le Français, il existe le corpus ANNODIS [*Péry-Woodley et al., 2009]*, composé de 86 documents provenant de l'Est Républicain et de Wikipédia, qui utilise 17 différentes classes pour qualifier les relations entre EDU.

Avant chaque analyse un pre-processing des données est fait: certains algorithmes de parsing nécessitent une structure d'arbre binaire *[Ji Eisenstein 2014]*, ils ont donc recours à des techniques de binarisation de d'arbres, comme le branchement à droite ou à gauche, etc. Certains choisissent de se limiter à plusieurs classes de relations *[Braud Denis 2013]*, et catégorisent donc certaines relations dans des classes plus généralistes afin limiter le choix et donc d'améliorer la classification faite en aval, ou choisissent simplement de ne pas analyser ceux qui n'utilisent pas ces relations.

Mais certains algorithmes d'apprentissage nécessitent une grande quantité de données d'apprentissage qui ne sont souvent pas disponibles. Une des causes est la difficulté à annoter ces données manuellement est que cela nécessite généralement l'assistance d'experts dans le domaine. Néanmoins, il existe des méthodes pour pallier à ce manque de données.

[Braud Denis 2013] explore différentes méthodes qu'il applique au Français. Il se concentre sur l'analyse de relations discursives implicites entre EDU (qui ne sont pas marqués explicitement par un connecteur discursif, comme *mais* pour marquer la relation d'opposition d'un EDU avec un autre). Le principe de *[Braud Denis 2013]* est de se baser sur des relations discursives explicites, qui sont plus faciles à détecter. Il est donc possible de générer automatiquement des corpus de relations explicites annotées, où l'on retire par la suite le connecteur discursif explicitant la relation entre les deux EDU en question. Mais cette stratégie est incertaine: il n'y a pas de garantie que les données générées automatiquement ressemblent aux données manuelles. L'auteur soulève des différences de distributions entre ces deux derniers: différence de répartition entre données relations inter-phrastiques (entre deux phrases différentes), et intra-phrastiques (relation entre EDU contenue dans une phrase), ou encore le fait que dans les données artificielles ne présentent par définition que des relations entre deux EDU contigus, alors que ce n'est pas le cas dans les données manuelles. Ces différences génèrent du bruit dans les données; il faut donc "guider" le modèle vers la distribution artificielle: c'est précisément la problématique que *[Braud Denis 2013]* traite.

En fonction de différents corpus utilisés les modèles prédictifs destinés à faire du *discourse parsing* peuvent varier fortement. On peut donc voir la nécessite d'assurer une cohérence globale entre données d'apprentissage.

II - MODÈLES

Voici une vue d'ensemble des modèles abordés dans les articles. Au début nous parlerons de modèles qui utilisent des traits, ou *features* plus ou moins surfaciques, pour ensuite se consacrer aux modèles dits *deep* qui se focalisent sur l'apprentissage d'une représentation lexicale d'unités élémentaire du discours.

Vous pourrez trouver quelques comparaisons rendant compte des différences entre structures de représentation du discours, ainsi que les algorithmes utilisés pour remplir la tâche d'analyse du discours. De plus, est parfois introduite la problématique à laquelle répond le modèle présenté.

A - Méthodes basées sur l'apprentissage à l'aide de features surfaciques

L'auteur dans *[Braud Denis 2013]* s'attache à classifier des différents types de relations discursives des couples d'EDU liées implicitement, et ne cherche pas à apprendre la structure discursive sous-jacente. Il a recours à un modèle de classification discriminant par régression logistique. Certains des traits utilisés sont énoncés: traits lexico-syntaxiques, indice de complexité syntaxique, nombre de syntagmes nominaux, verbaux, prépositionnels, adjectiviaux, adverbiaux, mais également des informations sur la tête d'un argument (par exemple : information temporelle, lemme d'éléments négatifs...), ou sur la paire d'arguments (par exemple : traits de position, indice de continuité thématique, ...).

D'autres part, face aux limitations de l'approche concurrente dite de "Constituency based discourse parsing" ayant une forte complexité en temps dû a la création de nombreux nœuds (mots et non EDU), les auteurs de *[Li Wang Cao Li 2014]* proposent une méthode d'analyse de dépendance structurelle moins coûteuse.

Ce modèle d'analyseur du discours consiste en la création d'arbres de dépendance structurelle binaires qui rendent compte des relations entre EDU dans un texte. Fortement inspirée de la "Rhetorical Structure Theory" *[Mann et Thompson 1988]*, cette approche se distingue dans sa structure de représentation du discours. En effet, contrairement aux arbres discursifs classiques (constituency based discourse parsing) où les feuilles représentent des EDU ou mots, et les noeuds internes des relations binaires entres EDU, les arbres de dépendances structurelles ont pour noeuds des EDU reliés par des arcs où l'étiquette correspond à la relation identifiée.

Pour pouvoir construire un spanning tree de dépendance optimal pour un document donné ou arbre couvrant le mieux le texte, la procédure consiste à appliquer un algorithme récursif de détermination d'arbres optimaux (Eisner ou Maximum spanning tree) couplé à au classifieur MIRA permettant de déterminer les relations dans l'arbre.

L'idée est d'abord de construire un arbre de dépendances avant d'annoter ses noeuds des relations qui lui sont associées. Pour se faire, l'algorithme d'Eisner explore l'espace de solutions des arbres de dépendances, pour en extraire le plus probable grâce à un classement de scores relatifs aux features et poids d'arcs. Le Maximum Spanning Tree Algorithm est une autre alternative; il permet d'en construire un récursivement. Le choix de l'algorithme à utiliser est fait selon la propriété projective ou non-projective de la structure du texte. En appliquant dans un deuxième temps le Margin Infused Relaxed Algorithm, un classifieur de structures multi-classe (de type statistique), on détermine alors le label de relation.

On remarque que la phase d'apprentissage n'est pas explicitée en détail dans l'article, il ne s'agit pas de l'objectif central de cet article.

B - Méthodes basées sur l'apprentissage de représentations

Ces méthodes tentent de trouver une représentation des relations entre différents EDU pour ensuite les comparer entre eux et extraire une structure discursive, ainsi que des labels pour les relations associées.

i - Réseau de neurones récurrents

L'article [Li Rumeng Li Hovy 2014] suggère l'usage de réseaux de neurones récurrents en tant que tentative de capturer une meilleure représentation des éléments de syntaxe et de sémantique des EDU, qui serait pertinents à l'analyse du discours.

Comme le rappellent les auteurs de cet article, un texte est dit cohérent lorsqu'il est possible de décrire sans ambiguïté le rôle que joue chaque unité du discours (EDU) vis-à-vis de ce dernier, et ce à tous les niveaux de granularité. Mais bien que l'on puisse obtenir un sens commun pour deux EDU partageant une relation et ayant chacun un sens individuel, il semblerait que l'analyse de textes entiers soit plus complexe que l'analyse de simples phrases.

En effet, il n'y a pas toujours une définition syntaxique pour les relations discursives et les membres d'une relation peuvent être placés individuellement dans des phrases ou paragraphes différents. Ces relations peuvent alors être identifiées par des marqueurs linguistiques (temps, usage de pronoms,conjonctions de coordination, etc). L'usage de réseaux de neurones récurrents est donc guidée par la confiance en une représentation plus complète qu'ils seraient capables de fournir.

L'analyseur présenté construit l'arbre discursif le plus vraisemblable pour un texte donné, selon le modèle appris sur la base RST-DT. Il suit une approche *bottom-up*. En effet, il s'agit à l'échelle des EDU, de décider si deux unités adjacentes sont connectées et si tel est le cas, par quelle relation; et de faire graduellement et récursivement une unification de sous arbres obtenus précédemment, jusqu'à couvrir totalement les phrases, puis le texte par un grand arbre de type RST.

La procédure de construction de l'arbre se fait à l'aide des deux classifieurs suivants:

- un classifieur binaire constitué de réseau de neurones récurrents déterminant la probabilité de fusionner deux EDU adjacents pour former un sous-arbre.

- un classifieur multi-classe également basé sur des réseaux de neurones récurrents pour sélectionner la relation discursive (ici la classe) la plus appropriée pour étiqueter le sous-arbre obtenu précédemment, et en calculer la représentation distribuée.

En ce qui concerne les algorithmes d'apprentissages utilisés, on peut distinguer les particularités suivantes:

- il y a deux fonctions de coût pour chaque réseau (binaire et multiclasse)
- une procédure de «backward propagation» standard (voir formule),
- présence de features supplémentaires (token en début ou fin d'un EDU, POS en début ou fin d'un EDU, deux EDU dans la même phrase),
- optimization: technique diagonale, variante de l'Adagrad très utilisé dans la communauté du *deep learning* (voir formule),
- les deux classifieurs (binaire & multi-classe) sont entrainés séparément sur RST-DT.

ii - Apprentissage de projections linéaires

L'article *[Ji Eisenstein 2014]* propose de représenter les différents EDU dans un espace latent dans lequel les relations entre EDU deviennent plus claires, et donc la structure discursive plus évidente. *[Ji Eisenstein 2014]* propose donc d'apprendre une transformation depuis une représentation 'Bag of words' vers la représentation d'un espace latent basé sur des features lexicaux, tout en apprenant à prédire des structures discursives basées sur cette même représentation latente.

Pour ce faire, l'auteur utilise le shift-reduce parsing: à chaque étape du parsing, l'on garde en mémoire une pile initialement vide, et une file, avec le premier EDU au début cette la file. Le parseur peut décider de faire l'action de *shift*, qui consiste à mettre le premier élément de la file dans la pile, ou l'action de *reduce*, où l'on affecte une relation ainsi qu'une hiérarchie (Noyau, Satellite) aux deux éléments de la file. Choisir une action appropriée peut donc être perçue comme un problème de classification, où l'on cherche l'action qui va maximiser une métrique prédéfinie.

La métrique choisie dans cet article est le produit scalaire entre le vecteur caractérisant l'action et la fonction de représentation des trois EDU mis en relation durant une étape du parsing. La fonction de représentation n'est rien d'autre qu'une projection linéaire de la représentation de trois EDU (représentation Bag-of-words) dans un espace latent.

L'apprentissage de la projection linéaire, ainsi que la représention des actions possibles est nécéssaire. L'auteur propose donc d'appliquer le Large-Margin Learning Framework. La projection linéaire est fixée durant l'étape de minimisation de la fonction cout (problème de type SVM), et à chaque itération, recalculée.

Ces actions de shift-reduce représentent l'arbre discursif. Chaque EDU n'a que pour paramètre sa représentation projetée dans l'espace latent. Afin de tirer le maximum d'informations des EDU, l'auteur introduit l'usage de d'autres features surfaciques, comme les mots aux extrémités des EDU, leur POS tagging, ou encore le Head word d'un EDU.

III - RÉSULTATS

Cette partie traite des résultats des expériences menées sur les données et des modèles présentés ci-dessus. Les expériences varient en fonction de la langue, du corpus, ainsi que des relations que l'on s'est données au préalable, une comparaison du score par rapport à un modèle serait parfois inexacte. On peut néanmoins constater une certaine cohérence au niveau de la manière d'exécuter l'expérimentation : les données sont toujours segmentées en données d'apprentissage et de test, et la méthode de validation croisée est systématiquement utilisée afin d'avoir un score le moins biaisé possible. Le score est toujours calculé à partir du f-score, qui est un compromis entre les notions de précision et de rappel. Les résultats sont ensuite comparés avec les résultats des modèles de l'état de l'art.

Pour *[Braud Denis 2013]*, qui est évalué sur le corpus *Annidis*, les meilleurs résultats sont de l'ordre de 45,6% d'exactitude, soit un gain de 5,9 % par rapport à un système n'utilisant que les données annotées manuellement .

Pour *[Ji Eisenstein 2014]*, l'analyse est faite est est confrontée à différents modèles de l'état de l'art, comme *[Hernault et al., 2010]*. Des résultats sont fournis pour différents paramètres du modèle, comme la dimension de l'espace latent représenté, ou la forme de la matrice de projection. Les résultats présentent une amélioration nette de 6 % sur les relations (donc de 71,13%), et de 2,5 % en nucléarité (donc de 61,75%) au score de F-mesure par rapport au modèles de l'état de l'art (HILDA) *[Hernault et al., 2010]*. Quant à la détection du span, on remarque une baisse de 1% (soit 81,60%). *[Ji Eisenstein 2014]* suppose que c'est sans doute dû aux traits syntaxiques et contextuels qui sont présents dans *[Hernault et al., 2010].* L'auteur de *[Ji Eisenstein 2014]* suggère qu'une combinaison des features de HILDA et des leurs pourraient encore parfaire leur modèle.

Pour *[Li Wang Cao Li 2014], l'*évaluation se fait sur le corpus RST-DT. Le système présenté dans cet article réalise de bonnes performances moyennes sur l'ensemble des mesures d'évaluation d'analyseurs discursifs. Les scores obtenus (avec grains moyens) sont de 82.9% ,73% et 60.6% sur les détections respectives de span, nucléarité et relation, soit 72.16% en moyenne, score qui est d'un bon ordre de grandeur face aux performances humaines de 77.3% (moyenne).

De plus, une autre métrique d'évaluation du modèle est la précision dans l'étiquetage des arcs sur les spann tree de dépendandes créés. On observe que les performance en appliquant l'algorithme MIRA sont supérieures à celles dans l'état de l'art (HILDA-manual, Fen, etc) avec 66.84% de précision.

Pour *[Li Rumeng Li Hovy 2014]*, l'évaluation se fait également sur le corpus RST-DT. La technique d'évaluation employée consiste en une comparaison de la structure et des labels de arbres structurels obtenus avec les structures et annotations «gold-standard» de la base de tests. La formule d'évaluation est la suivante : diviser le nombre de similarités dans les constituants des arbres par le nombre de constituants total.

Pour compléter la phase d'évaluation, les auteurs décident de comparer les performances de leur modèle avec d'autres analyseurs de discours concurrents que sont le HILDA, celui de Joty et Al, de Feng et Hirst, ou encore de Ji et Eisenstein.

Contrairement aux modèles concurrents cités précédemment réalisant des scores optimaux sur quelques mesures d'évaluation , avec 85.7% de détection de span pour Feng et Hirst, 71.1% et 61.6 de détection de nucléarité et de relations pour Ji et Eisenstein, le modèle présenté dans cet article réalise un des meilleurs scores en terme de moyenne sur l'ensemble des mesures d'évaluation, sans toutefois atteindre un score optimal sur une seule mesure. En effet, les scores obtenus (avec features) sont de 84.0% ,70.8% et 58.6% sur les détections respectives de span, nucléarité et relation, soit 71.13% en moyenne, score relativement bon face aux performances humaines de 77.3% (moyenne).

De plus, les auteurs attirent notre attention sur le fait qu'il faut prendre en considération différents types de relation dans le traitement de l'information (convolution) pour éviter de voir les performances s'effondrer.

CONCLUSION :

Nous avons pu voir grâce aux résultats des expériences que l'analyse discursive est une tâche encore difficile du traitement automatique du langage. Les causes de ce score relativement faible sont nombreuses et variées: les données annotées ne sont pas toujours présentes en quantité suffisante, il existe un grand nombre de lexicalisations alternatives possibles (on retrouve même des lexicalisations différentes d'un même extrait de corpus en fonction de deux experts différents dans les corpus annotés), etc.

Nous avons aussi pu confronter différents modèles, ainsi que leurs traits utilisés, que nous avons séparés en deux catégories: les modèles d'apprentissage uniquement basés sur des traits lexicaux syntaxiques et surfaciques, ainsi que des modèles qui apprennent une représentation, pour tenter de capturer des traits sémantiques qui caractérisent au mieux les différentes relations entre EDU. Comme nous avons pu le voir précédemment, il y a une tendance actuelle vers l'utilisation de modèles basés sur l'apprentissage de représentations.

Une explication plausible de cette tendance est le fait que les traits de surface et de syntaxe ne parviennent pas à expliquer correctement ce que constitue réellement une distinction sémantique, il faut donc appliquer ces modèles dits *profonds*. Les résultats semblent en attester l'utilité : un des modèles deep faisant preuve des meuilleures performances dans le *discourse parsing* est le modèle DPLP *[Ji Eisenstein 2014]*, avec une amélioration nette de 6 % sur le score de classification les relations, et de 2,5 % en nucléarité au score de F-mesure par rapport aux précédents modèles de l'état de l'art(2010 à 2013).

RÉFÉRENCES:

[Braud Denis 2013] Identification automatique des relations discursives «implicites» à partir de données annotées et de corpus bruts, 2013, **Chloé Braud**, ALPAGE, INRIA Paris-Rocquencourt & Université Paris Diderot, **Pascal Denis**, MAGNET, INRIA Lille Nord-Europe.

[Ji Eisenstein 2014] Representation Learning for Text-level Discourse Parsing, 2014, **Yangfeng Ji**, School of Interactive Computing , Georgia Institute of Technology INRIA, and **Jacob Eisenstein**, School of Interactive Computing, Georgia Institute of Technology.

[Li Wang Cao Li 2014] Text-level Discourse Dependency Parsing, **Sujian Li**, **Liang Wang**, **Ziqiang Cao**, et **Wenjie Li**,2014

[Li Rumeng Li Hovy 2014] Recursive Deep Models for Discourse Parsing, **Jiwei Li, Rumeng ,Li** et **Eduard Hovy**, 2014

[NLPERS] http://nlpers.blogspot.fr/2010/08/why-discourse-structure.html

[Mann et Thompson, 1988] **Mann, W. C.** et **Thompson**, **S. A**. (1988). Rhetorical structure theory : Toward a functional
theory of text organization. Text, 8(3):243–281.

[Asher et Lascarides, 2003] **Asher, N.** et **Lascarides**, **A**. (2003). Logics of conversation. Cambridge University Press.

[Daniel Marcu, Magdalena Romera, and Estibaliz Amorrortu (1999)] [**Daniel Marcu, Magdalena Romera**, and **Estibaliz Amorrortu** (1999)]. Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. *The Workshop on Levels of Representation in Discourse,* pages 71-78, Edinburgh, Scotland, July 1999.

[Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski (2002)] **Lynn Carlson**, **Daniel Marcu**, and **Mary Ellen Okurowski** (2002). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory.

[Prasad et al., 2008] **Prasad**, **R.,Dinesh**, **N.,Lee**, **A.,Miltsakaki**, **E.,Robaldo**, **L.,Joshi, A.** et **Webber, B**. (2008).
The penn discourse treebank 2.0. In Proceedings of LREC, page 2961.

[Péry-Woodley et al., 2009] **Péry-Woodley**, **M.P.,Asher**, **N.,Enjalbert**, **P.,Benamara**, **F.,Bras**, **M.,Fabre**, **C.,Ferrari**, **S.,Ho-Dac**, **L.M.,Ledraoulec**, **A**. et **Mathet**, **Y**. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. Actes de TALN 2009.

[Hernault et al., 2010] **Hugo Hernault**, **Helmut Prendinger**, **David A. duVerle**,
and **Mitsuru Ishizuka**. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification.
Dialogue and Discourse, 1(3):1–33.