Identifier les relations discursives implicites en combinant données naturelles et données artificielles

Chloé Braud, Pascal Denis

Identifier les relations discursives implicites en combinant données naturelles et données artificielles

Chloé Braud* — **Pascal Denis****

* ALPAGE, *Université Paris Diderot & INRIA Paris-Rocquencourt*
chloe.braud@inria.fr

** MAGNET, *INRIA Lille Nord Europe*
pascal.denis@inria.fr

RÉSUMÉ. *Cet article présente les premières expériences sur le français d'identification automatique des relations discursives implicites (i.e., non marquées par un connecteur). Nos systèmes exploitent des exemples implicites annotés, ainsi que des exemples implicites artificiels obtenus à partir d'exemples explicites par suppression du connecteur, une méthode introduite par Marcu et Echihabi (2002). Les précédentes études sur l'anglais montrent que l'utilisation à l'entraînement des données artificielles dégrade largement les performances sur les données naturelles, ce qui reflète des différences importantes en termes de distribution. Ce constat, qui tient aussi pour le français, nous a amenés à envisager différentes méthodes, inspirées de l'adaptation de domaine, visant à combiner plus efficacement les données. Nous évaluons ces méthodes sur le corpus* ANNODIS *: notre meilleur système obtient* 41,7 % *d'exactitude, soit un gain significatif de* 4,4 % *par rapport à un modèle n'utilisant que les données naturelles.*

ABSTRACT. *This paper presents the first experiments on French in automatic identification of implicit discourse relations (i.e. relations that lack an overt connective). Our systems exploit hand-labeled implicit examples, along with artificial implicit examples obtained from explicit examples by suppressing their connective, following Marcu et Echihabi (2002). Previous work on English shows that using artificial data for training largely degrades performance on natural data, reflecting important differences in the distribution. This conclusion, that also holds for French, has led us to consider various methods inspired by domain adaptation to better combine the data. We evaluate these methods on the* ANNODIS *corpus: our best system achieves a* 41.7 % *accuracy, that is a significant gain of* 4.4 % *compared to a model using only the natural data.*

MOTS-CLÉS : *structure discursive, relations discursives implicites, apprentissage automatique.*

KEYWORDS: *discourse structure, implicit discourse relations, machine learning.*

1. Introduction

L'analyse discursive a pour but de rendre compte de la cohérence d'un texte, cohérence reposant sur le fait que les propositions ne s'interprètent pas isolément les unes des autres, elles sont liées les unes aux autres par des relations dites discursives ou rhétoriques (type *explication, contraste...*) de manière à former une structure globale couvrant tout le document. Ainsi, dans l'exemple 1[1] les propositions formant la première phrase sont liées par une relation de *contraste*, et la phrase entière est elle-même liée à la phrase suivante par une relation de *continuation*.

(1) { { [La hulotte est un rapace nocturne]$_1$ [mais elle peut vivre le jour.]$_2$ }$_{contraste}$
[La hulotte mesure une quarantaine de centimètres.]$_3$ }$_{continuation}$
(ANNODIS - document WK_-_hulotte)

Différents cadres théoriques et formalismes ont été développés pour l'analyse discursive, notamment la *Rhetorical Structure Theory* (RST) (Mann et Thompson, 1988), la *Segmented Discourse Representation Theory* (SDRT) (Asher et Lascarides, 2003) et le formalisme des *Discourse Lexicalized Tree Adjoining Grammars* (D-LTAG) (Webber, 2004). Ces cadres définissent tous un jeu de relations différent, avec certains points communs (Roze, 2013). Les relations peuvent lier des segments discursifs de type élémentaire (EDU), en général des propositions (comme les segments 1, 2 et 3 dans l'exemple 1). Elles peuvent aussi s'établir entre des segments dits complexes, c'est-à-dire des paires de segments, élémentaires ou complexes, eux-mêmes liés par une relation (comme la paire de segments $(1, 2)$ liés par la relation *contraste* dans l'exemple 1). Enfin, les structures finales obtenues sont différentes : des arbres en RST et des graphes en SDRT.

Ces cadres et formalismes ont permis l'annotation de corpus. Pour la langue anglaise, on peut citer le corpus *RST Discourse TreeBank* (RST DT) suivant la RST (Carlson *et al.*, 2001) et le *Penn Discourse TreeBank* (PDTB) suivant le formalisme D-LTAG (Prasad *et al.*, 2008). Pour la langue française, le corpus ANNODIS (Afantenos *et al.*, 2012) a été développé dans le cadre de la SDRT. Les cadres qui sous-tendent ces corpus impliquent des différences au niveau des structures annotées et des ensembles de relations. Concernant les jeux de relations : le RST DT comporte 78 relations regroupées en 16 classes, ANNODIS 17 relations et le PDTB propose une organisation hiérarchique avec 4 classes, 16 types et 23 sous-types.

Le PDTB présente une autre caractéristique : une distinction explicite est établie à l'annotation entre différents types de relations. Ce corpus est fondé sur le formalisme D-LTAG, l'idée étant que certains éléments lexicaux spécifiques, appartenant à une

1. On considère comme un exemple de relation une paire de segments textuels (ou arguments) associée à une relation (ou type de relation, comme *explication, résultat...*). Dans les exemples cités, on adopte la convention suivante : les connecteurs sont inclus dans l'argument constitué par la clause dont ils sont syntaxiquement dépendants. Les crochets délimitent les segments élémentaires et les accolades délimitent les segments complexes.

liste fermée ont pour fonction de déclencher une relation discursive : les connecteurs[2] ou marqueurs discursifs (*mais, parce que, pour...*). L'annotation part d'une liste de connecteurs préétablie qui sont repérés et distingués des formes en emploi non discursif (comme *pour* marquant une attribution). Leurs arguments, vus comme des EDU, et la ou les relations qu'ils déclenchent sont ensuite annotés. Ainsi dans l'exemple 2a, *pour* déclenche un *but* entre les segments 1 et 2. On parle de relation *explicite* car explicitement marquée par un connecteur discursif. Par opposition, lorsque aucun connecteur n'est présent, on parle de relation *implicite* ou non marquée, comme l'exemple de relation d'*explication* entre les segments 1 et 2 dans l'exemple 2b[3].

(2) a. { [Dans un premier temps, le président nancéien invita Pablo à cesser sa carrière]$_1$ [**pour** intégrer le staff technique.]$_2$ }$_{but}$

b. { [Hier en fin d'après-midi, Pablo a signé un nouveau contrat de deux ans,]$_1$ [acceptant finalement les propositions de Jacques Rousselot.]$_2$ }$_{explication}$
(ANNODIS - document Extr_estrep-025_PabloCorrea)

Ces corpus ont permis le développement de systèmes d'analyse automatique complets ou limités aux différentes sous-tâches. Bien que variables selon le cadre théorique sous-jacent, les étapes des systèmes complets incluent généralement une étape de segmentation correspondant à l'identification des EDU, une étape d'attachement permettant de décider de la présence ou de l'absence d'une relation entre deux segments, une étape d'identification de la relation liant deux EDU et enfin un processus de construction de la structure complète. Les systèmes complets obtiennent encore des performances modestes, ainsi Sagae (2009) et Hernault *et al.* (2010) obtiennent respectivement 44,5 % et 47,3 % de F-mesure sur le RST DT et Lin *et al.* (2010) rapportent un score de 33,0 % de F-mesure sur le PDTB.

Dans ces systèmes complets, l'étape d'identification d'une relation est cruciale, car une relation peut contraindre la structure (*via* la contrainte de la frontière droite en SDRT par exemple) ou le déclenchement d'autres relations (Roze, 2013). De plus un système identifiant automatiquement les relations pourrait être utile à d'autres systèmes. Pour l'analyse temporelle, le fait d'identifier une relation d'*explication* dans l'exemple 3 entre les segments 1 et 2 permet d'inférer le bon ordre temporel entre les événements, « pousser » avant « tomber ». Pour un système de résumé automatique, l'identification d'une relation entre deux segments peut permettre de déterminer si l'un des segments est supprimable, par le principe de hiérarchie entre les segments défini par exemple en SDRT sous le terme de relations coordonnantes ou subordonnantes.

2. On adopte ici la définition des connecteurs correspondant au lexique utilisé décrit en section 5.

3. Ici, « finalement » peut bien être vu comme un connecteur, mais, selon notre lexique, il ne déclenche pas la relation *explication*. La présence du participe présent « acceptant » peut être considérée comme une marque, ambiguë, de la relation d'*explication*. On considère donc ici comme implicite les cas où aucun connecteur de notre lexique n'est présent pour lexicaliser la relation annotée.

(3) { [Paul est tombé.]$_1$ [Marie l'a poussé.]$_2$ }$_{explication}$
(Exemple construit)

Différentes études ont cherché ces dernières années à améliorer les performances sur cette tâche particulière d'identification d'une relation entre deux segments textuels. Sur le RST DT, Hernault *et al.* (2010) rapportent 66,8 % d'exactitude. Sur ANNODIS en français, Muller *et al.* (2012) rapportent 44,8 % d'exactitude. Les études sur le PDTB ont séparé le problème en deux sous-tâches : l'identification des exemples de relations explicites et de relations non explicites. Ces études ont montré que l'identification d'une relation dans le cas explicite est beaucoup plus facile, les systèmes aboutissent à des performances de l'ordre de 94 % d'exactitude pour les 4 grandes classes de relations du PDTB (Pitler et Nenkova, 2009) et de 86 % pour les 16 relations de niveau 2 (Lin *et al.*, 2010) et ce, même en ne prenant que le connecteur comme trait. En revanche, dans le cas non explicite, les meilleures performances pour les 4 classes sont de l'ordre de 65,54 % (Pitler *et al.*, 2009) quand seules les données manuellement annotées sont utilisées pour l'apprentissage, et de 72,2 % (Wang *et al.*, 2012) quand des données explicites sont ajoutées. Pour 11 relations de niveau 2, Lin *et al.* (2009) rapportent 40,2 % d'exactitude sur les données manuellement annotées. Les performances sont donc encore modestes quand une relation n'est pas explicitement marquée, ce qui correspond à environ la moitié des données du PDTB. Dans ce cas, les indices sont complexes, et les performances basses sont probablement dues à un manque de données annotées.

Pour faire face à ce manque de données, Marcu et Echihabi (2002) ont proposé une méthode simple visant à créer automatiquement de nouveaux exemples de relations implicites à partir de données brutes. Cette méthode repose sur une détection des segments et un repèrage automatique des connecteurs, idéalement non ambigus (ou désambiguïsés) en emploi et en relation : ces connecteurs sont alors supprimés[4] et utilisés comme annotation de la relation identifiée. Outre sa simplicité, cette approche a aussi l'attrait de permettre la création d'un grand nombre d'exemples, puisqu'elle opère sur des données brutes. En contrepartie, cette méthode peut potentiellement pâtir d'erreurs faites lors de la segmentation et de la détection des connecteurs. Elle fait, par ailleurs, plusieurs hypothèses importantes. D'une part, elle postule que le connecteur est redondant avec son contexte : il faut en effet qu'il reste suffisamment d'information, une fois le connecteur supprimé, pour qu'on puisse encore identifier la relation. En outre, elle suppose que les données construites automatiquement (ou artificielles) suivent la même distribution que les données manuellement annotées (ou naturelles). Nous verrons en détail dans les sections qui suivent que ces hypothèses ne sont pas toujours vérifiées.

D'ailleurs, les précédentes études reposant sur cette approche ne sont le plus souvent pas parvenues à une amélioration des performances quand ces nouvelles données sont utilisées comme seules données d'entraînement. La baisse de performance observée, par exemple dans (Sporleder et Lascarides, 2008), montre que les nouvelles

4. Des détails sur cette méthode sont donnés en section 3.

données ne permettent pas une bonne généralisation aux exemples de relations implicites manuellement annotés. Dans leur étude, Wang *et al.* (2012) mettent en place un principe de sélection parmi les exemples de relations originellement explicites, issus du PDTB, afin de sélectionner les exemples ressemblant aux exemples naturels de relations implicites servant d'évaluation. Cette méthode leur permet de gérer les différences en termes de distribution et d'améliorer les performances, notamment en termes de micro-exactitude en multiclasse, par rapport à un système n'utilisant que les données manuellement annotées. Contrairement à eux, nous ne disposons pas d'un corpus dans lequel les connecteurs ont été explicitement annotés. De plus, comme nous disposons de moins de données naturelles, l'utilisation de données brutes nous permet d'acquérir un très grand nombre de données artificielles, ce qui pourrait potentiellement compenser le faible volume de données manuellement annotées plus crucial encore en français (252 exemples utilisés ici contre plus de 14 000 pour le PDTB) mais qui introduit du bruit absent de leur étude. Enfin, leur méthodologie nécessite l'écriture de règles manuelles pour la phase de sélection des exemples et exclut des exemples naturels des données d'entraînement ce qui pourrait limiter la stratégie.

Construire un modèle sur des données de distribution sous-jacente différente de celle des données sur lesquelles on va l'utiliser va à l'encontre des hypothèses généralement requises par un système d'apprentissage et risque d'aboutir à des performances basses. Cette configuration ressemble à celle étudiée dans le cadre de l'adaptation de domaine où le problème consiste à adapter un classifieur à des données dites cibles (« out-of-domain ») issues d'un domaine différent de celui des données manuellement annotées disponibles, les données sources (« in-domain ») [5]. Nous avons donc un problème de base similaire, ce qui nous a amenés à mettre en œuvre des stratégies inspirées de ce cadre, méthodes combinant les données ou les modèles. Grâce à des méthodes simples, nous avons obtenu des améliorations significatives par rapport à un simple entraînement sur les données naturelles disponibles.

Dans cet article, nous présentons d'abord les études précédentes dans lesquelles ont été utilisées des données annotées automatiquement à partir des connecteurs (section 2). Nous étudions ensuite les problèmes posés par cette méthode (section 3) avant de présenter la stratégie envisagée pour répondre à ces problèmes (section 4). Nous présentons enfin les données (section 5) puis les expériences réalisées et les résultats obtenus (section 6). Enfin nous présenterons les perspectives futures (section 7).

5. La notion de « domaine » est à comprendre dans un sens large. Elle distingue généralement des textes portant sur des thèmes différents (économique *vs* sportif) ou issus d'un même support mais couvrant des périodes temporelles différentes (pour des articles de journaux notamment). Plus formellement, dire que l'on dispose de données dans un domaine, c'est dire que l'on dispose de données suivant une certaine distribution. Les données « out-of-domain » sont donc des données suivant une autre distribution (Daumé III et Marcu, 2006).

2. Études précédentes

Nous présentons dans cette section les précédentes études ayant utilisé un corpus de données annotées automatiquement à partir d'exemples explicites. On note qu'une comparaison directe entre ces études n'est pas possible, d'abord parce que le corpus annoté manuellement diffère, RST DT ou PDTB, ensuite parce que, même en utilisant le même corpus, les configurations diffèrent également. Ainsi, le corpus RST DT ne sépare pas explicitement les données explicites et implicites, et les études sur ce corpus, sans préciser en général comment les données ont été séparées, recensent des quantités différentes donc utilisent des corpus d'évaluation différents. Par exemple, pour *contrast*[6], Marcu et Echihabi (2002) relèvent 177 exemples implicites. En revanche, Wang *et al.* (2012) ne considèrent que les occurrences de relations interphrastiques et en recensent 889 (dont 311 de test). Enfin, Sporleder et Lascarides (2008) sélectionnent un échantillon aléatoire et utilisent 238 exemples pour l'évaluation. Si le PDTB devrait amener plus de consensus, les auteurs utilisent des ensembles d'entraînement et d'évaluation différents. La première raison tient à la définition de « implicite » : tandis que Wang *et al.* (2012) restreignent leur étude aux exemples annotés comme implicites dans le PDTB, Pitler *et al.* (2009) et Park et Cardie (2012) incluent les exemples annotés comme *Entity* et *AltLex*, formant un ensemble d'exemples non explicites. Ensuite, une paire de segments peut être annotée avec plusieurs relations dans le PDTB. Certains auteurs les prennent toutes en compte dans leurs systèmes en dupliquant les exemples pour chaque annotation (Pitler *et al.*, 2009). Au contraire, Wang *et al.* (2012) ne prennent en compte que la première relation annotée. Enfin, les corpus d'entraînement diffèrent également à cause des rééchantillonnages aléatoires (« down-sampling ») effectués pour les classifieurs binaires, et éventuellement d'une optimisation de ces corpus (Park et Cardie, 2012).

Marcu et Echihabi (2002) ont les premiers proposé d'utiliser des exemples explicites pour la construction d'un système d'identification des relations implicites. Les connecteurs étant peu ambigus, ils ne déclenchent en général qu'une seule relation, il est relativement facile d'extraire des exemples explicites de données brutes et de leur assigner la bonne relation. Ces auteurs ont utilisé deux corpus bruts anglais (près de 48 millions de phrases) et ont défini des motifs fondés sur la position du connecteur et la ponctuation pour extraire des exemples de 4 relations (*contrast, cause-explanation-evidence, condition, elaboration*). Ainsi, ils extraient des exemples de *contrast* à partir de *but* si le connecteur débute une phrase, les arguments sont alors la phrase contenant *but* et la phrase précédente. Si *but* apparaît à l'intérieur d'une phrase, les arguments sont les segments délimités par le connecteur. Les auteurs extraient aussi des exemples correspondant potentiellement à une absence de relation en se fondant sur des phrases éloignées (à l'intérieur d'un même document ou entre documents). Les données sont représentées par le produit cartésien des mots sur les arguments, l'idée

6. Lorsque nous faisons référence à de précédentes études, nous conservons les notations utilisées dans ces éudes pour le nom des relations car, selon le cadre théorique sous-jacent, un même nom ne recouvre pas nécessairement la même définition.

étant d'identifier des paires de mots dont le lien correspondrait à une relation, comme une opposition « embargo/legally » pour des contrastes. Les auteurs testent sur des exemples artificiels des classifieurs binaires de type bayésien naïf : ceux-ci dépassent tous les 50 % d'exactitude. Ils testent également sur les mêmes données un classifieur bayésien naïf multiclasse (4 relations discursives et 2 classes marquant l'absence de relation inter ou intradocument) dont l'exactitude est également de l'ordre de 50 %. Il est important de noter que Marcu et Echihabi (2002) n'ont pas vraiment testé leur approche sur des données implicites naturelles, leur ensemble d'évaluation naturel contient aussi des explicites et d'ailleurs ils conservent le connecteur dans les données extraites des corpus bruts. Ils estiment de manière très informelle les résultats potentiels de leur système sur les exemples de relations implicites naturels. Cela dit, les performances obtenues sur les données artificielles sont largement supérieures au hasard, ce qui accrédite l'hypothèse de redondance du connecteur : pour au moins une partie de ces données, la relation reste identifiable une fois le connecteur supprimé. À la suite de Marcu et Echihabi (2002), différents auteurs ont repris le principe de données annotées automatiquement et testé cette méthode de différentes façons.

Blair-Goldensohn *et al.* (2007) ont testé la méthode à la fois sur un corpus brut et sur le corpus du PDTB avec 2 relations (*contrast, cause-explanation-evidence*) et une classe représentant l'absence de relation (intradocument) et des classifieurs binaires. Ils ont cherché à limiter le bruit dans les données notamment grâce à des pré-traitements. Leurs résultats sont difficiles à comparer puisque les données et la configuration sont différentes. Surtout, ils s'évaluent sur des données artificielles (5 000 exemples par relation) et sur des données implicites naturelles du PDTB (en sélectionnant aléatoirement 280 exemples par relation) mais ne se comparent pas à un entraînement sur des données naturelles. Leurs expériences ne suffisent pas pour savoir si les données artificielles mènent à une amélioration de l'identification des exemples de relations implicites naturels, mais seulement que ces données synthétiques permettent d'aboutir à des performances supérieures au hasard sur les données naturelles, pour les relations considérées et en classification binaire. De plus, les prétraitements utilisés n'améliorent pas les performances en termes d'exactitude sur l'ensemble de test naturel.

Pitler *et al.* (2009) construisent un système standard d'identification des relations implicites sur le PDTB, avec un entraînement et un test sur des données manuellement annotées avec des traits repris d'études précédentes. Ils utilisent les 4 relations de premier niveau (*temporal, contingency, comparison, expansion*) pour lesquelles ils construisent des classifieurs binaires. Leur utilisation de données artificielles (corpus de (Blair-Goldensohn *et al.*, 2007) ou données explicites du PDTB) se limitent à l'extraction dans ces données de paires de mots (produit cartésien sur les arguments), reprenant l'idée de base de Marcu et Echihabi (2002) dont ils cherchent à tester la validité. Leur étude montre, pour les deux relations considérées dans (Blair-Goldensohn *et al.*, 2007), que l'utilisation de ces paires de mots dégrade généralement les performances, du moins l'utilisation de ces seuls traits conduit aux systèmes parmi les plus mauvais mais les auteurs ne présentent pas de système combinant ces traits à d'autres. De plus les auteurs montrent que les paires de mots présentant le plus haut

gain d'information ne sont pas des paires sémantiquement liées, mais des paires de mots fonctionnels. Cependant, l'étude montre aussi une combinaison potentiellement efficace entre données naturelles et artificielles : si seules les paires de mots extraites des données artificielles correspondant à un gain d'information non nul sur les données implicites sont conservées, on obtient le meilleur système pour *comparison* et le deuxième meilleur système pour *contingency*. Ces systèmes ont cependant été dépassés depuis par Park et Cardie (2012) avec des modèles plus simples qui n'utilisent pas de données artificielles.

Sporleder et Lascarides (2008) ont mené une étude approfondie sur le principe même de cette stratégie et ont construit un système utilisant des traits linguistiquement motivés. Ils testent cette stratégie sur des données du corpus RST DT en choisissant 5 relations de la SDRT (en établissant une correspondance entre les relations). Ils concluent sur l'impossibilité d'utiliser directement ces données en entraînement et posent différentes questions remettant en cause le principe même de cette stratégie. Nous reprenons ces questions dans la section suivante. Ces auteurs ont montré qu'un entraînement sur un petit ensemble de données annotées manuellement aboutissait à des performances largement supérieures (1 051 exemples, exactitude de 40,3 %) à un entraînement sur une grande quantité de données annotées automatiquement à partir d'exemples explicites (72 000 exemples, exactitude de 25,8 %). Cette baisse de performance reflète la différence de distribution entre les deux ensembles de données, mais les résultats obtenus sont encore supérieurs au hasard (fixé à 20 %).

Wang *et al.* (2012) ont utilisé un principe de sélection des exemples d'entraînement fondé sur un algorithme non supervisé. Le système est testé sur les 4 classes du PDTB (*temporal, contingency, comparison, expansion*) et sur le RST DT (*temporal, contrast, cause, background*). Des règles manuelles permettent d'identifier parmi les exemples d'entraînement un petit ensemble « graine » d'exemples considérés comme « typiques » d'une relation. Par exemple, pour *comparison* et *contingency*, une paire de mots de polarité opposée doit être identifiée dans les arguments. Pour *temporal*, l'un des arguments doit contenir les mots « *including, following, last, first, second* ». Un algorithme de clustering étend ensuite l'ensemble « graine » et un algorithme supervisé est utilisé sur l'ensemble d'entraînement ainsi constitué. Pour le RST DT les auteurs disposent d'environ 2 400 exemples artificiels et 3 900 implicites naturels. Pour le PDTB, le corpus artificiel correspond aux exemples explicites du corpus. Les données sont représentées par des attributs binaires repris des études précédentes (traits lexicaux et sémantiques). En multiclasse, les auteurs rapportent 51,3 % d'exactitude avec le classifieur bayésien naïf (NB) et 53,5 % avec des arbres de décision (DT) quand toutes les données implicites naturelles sont utilisées. Quand les données artificielles sont utilisées seules, les performances tombent à 34,1 % (NB) et 42,6 % (DT). L'union des données permet d'améliorer ces performances : 42,3 % (NB) et 51,4 % (DT). Enfin la méthode de sélection d'exemples introduite permet une amélioration importante : 68,3 % (NB) et 72,2 % (DT). Dans cette configuration, le nombre d'exemples artificiels et naturels est similaire (respectivement 6 753 et 7 816). On note cependant que les classifieurs binaires sur le PDTB obtiennent des scores de F-mesure plus bas que ceux rapportés dans l'étude de Park et Cardie (2012) n'utilisant pas de

données artificielles mais des prétraitements et des combinaisons de traits différentes, cependant comme noté en introduction une comparaison directe entre ces études n'est pas possible. Une comparaison stricte de ces deux stratégies et une estimation de leur coût permettraient de déterminer la plus efficace. Enfin Wang *et al.* (2012) notent que les exemples correctement identifiés par leur système correspondent en général à des exemples « typiques », il n'est donc pas clair que ce système puisse aboutir à des performances encore supérieures, bien qu'une étude sur la représentation des données et son influence sur la typicité des exemples reste à mener.

3. Problématique

L'utilisation d'exemples artificiels pour l'apprentissage des relations implicites repose sur différentes hypothèses que la plupart des précédentes études n'ont souvent pas explicitées clairement. Comme nous l'avons déjà évoqué, cette méthode suppose tout d'abord une certaine redondance du connecteur avec son contexte : il doit rester suffisamment d'information, une fois le connecteur supprimé, pour pouvoir identifier la relation. Ensuite, il faut aussi que les exemples artificiels suivent une distribution suffisamment proche (notamment en termes des étiquettes ou de l'association entre paires de segments et étiquettes) de celle des exemples naturels pour pouvoir espérer que les généralisations obtenues par le système de classification puissent s'appliquer aux exemples naturels et mener à des taux d'erreur réduits sur ces données. La suite de cette section revient en détail sur ces deux hypothèses.

3.1. *Redondance du connecteur*

Pour rappel, la méthode d'annotation automatique se fonde sur des exemples bruts identifiés comme explicites : on supprime le connecteur et on annote la relation que le connecteur lexicalisait. À titre d'illustration, considérons l'exemple 4 de *contraste* formé des segments 1 et 2 contenant le connecteur *cela dit* : l'exemple d'une relation implicite synthétique associé est un exemple de *contraste* formé des segments 1 et $2'$.

(4) { [Elle était très comique, très drôle.]$_1$ [(***Cela dit***) [, le drame n'était jamais loin.]$_{2'}$]$_2$ }$_{contraste}$

(*Est Républicain* - Exemple artificiel)

De manière à mieux évaluer l'hypothèse de redondance, nous envisageons les différents effets discursifs de la suppression d'un connecteur. Schématiquement, on identifie trois cas possibles :

– la suppression du connecteur a peu ou *pas d'effet* : la relation reste inférable en son absence, le connecteur est redondant avec son contexte ;

– la suppression du connecteur rend le *discours incohérent* : le discours devient innacceptable et ce non pas simplement à cause d'effets syntaxiques ;

– la suppression du connecteur modifie la *relation inférée* : la relation lexicalisée par le connecteur n'est plus inférable mais une autre relation est inférée.

3.1.1. *Connecteur redondant*

Selon Moeschler (2002), un connecteur « donne des instructions sur la manière de relier [des] unités » et « impose de tirer de la connexion discursive des conclusions qui ne seraient pas tirées en [son] absence ». Cependant, comment expliquer alors que le même lien, temporel en 5a ou causal en 5b, soit inféré en présence ou en absence des connecteurs (respectivement *et* et *parce que*) ? Cette possibilité semble dépendre essentiellement du contenu des segments liés et de préférences sur les relations inférables en l'absence de marques explicites.

(5) a. { [L'avion atterrit] [(***et***) les passagers descendirent.] }$_{narration}$

b. { [Paul tomba,] [(***parce que***) Marie l'avait poussé.] }$_{explanation}$

((Moeschler, 2002) - Exemple construit)

Dans leur étude, Soria et Ferrari (1998) demandent à des sujets d'identifier une relation (parmi *additive*, *consequential* et *contrastive*) dans des exemples où le connecteur est ou non supprimé (l'exemple restant grammatical). L'étude montre que l'identification est effectivement plus difficile sans connecteur (on passe de 72,6-88,9 % à 42,7-64,3 % d'identification correcte par relation), mais le statut des relations diffère : la relation *additive* semble prédite par défaut (car c'est celle qui est prédite par erreur le plus souvent), et la relation *consequential* reste, quant à elle, bien prédite reflétant une préférence des locuteurs pour ce type de relations. Ces conclusions correspondent à celles de Sanders (2005) : les lecteurs cherchent à inférer la relation la plus informative, un lien causal, puis, si le contenu des segments ne permet pas d'inférer une relation plus spécifique, un simple lien de continuité temporelle est inféré. On retrouve ces conclusions dans une étude menée sur le PDTB par Asr et Demberg (2012) : les exemples mettant en jeu un enchaînement temporel non linéaire sont plus souvent marqués que les autres et les relations causales sont plus facilement implicites. Cela ne signifie pas que seules les relations temporelles et causales restent identifiables en l'absence de connecteur, mais leur inférence est préférée. On note par ailleurs que l'entraînement et l'évaluation d'un modèle sur les données artificielles aboutissent à des performances supérieures au hasard pour les relations considérées qui restent donc identifiables dans une partie des données.

3.1.2. *Cohérence du discours*

Dans certains cas en revanche, le connecteur est nécessaire : sa suppression peut rendre le discours incohérent, c'est un effet noté par Asher et Lascarides (2003) dans le cas des contrastes de type violation d'attente. Ainsi dans l'exemple donné par ces auteurs et repris en 6a, la suppression du connecteur *but* conduit à un discours bizarre jugé incohérent. Cependant, dans les contrastes formels, où il n'y a pas de relation logique entre les conditions de vérité des arguments mais plutôt un contraste dû à

une différence de contenu, le marqueur n'est pas nécessaire : dans l'exemple 6b, la suppression du connecteur aboutit à une perte d'information mais le discours reste acceptable.

(6) a. { [John likes sports.] [(***But***) he hates football.] }*violation of expectation*

b. { [John has green eyes.] [(***But***) Mary has blue eyes.] }*formal contrast*

((Asher et Lascarides, 2003) - Exemple construit)

3.1.3. *Modification de la relation inférée*

Enfin, la suppression du connecteur peut modifier le type de relation inférable entre deux segments. Cet effet rejoint la question des exemples de relations implicites d'interprétation indéterminée dans (Moeschler, 2002). Concernant cet effet, Sporleder et Lascarides (2008) avaient proposé l'exemple 7 attesté ayant servi à produire un exemple artificiel : la suppression du connecteur *although* marquant un *contrast* annule l'inférence de cette relation mais le discours reste cohérent, on identifie une relation de type *continuation*. Ici les deux relations étaient présentes à l'origine, mais la suppression du connecteur ne rend plus possible que l'inférence de la relation implicite de *continuation*.

(7) { [(***Although***) the electronics industry has changed greatly,] [possibly the greatest change is that very little component level manufacture is done in this country.] }*contrast, continuation*

((Sporleder et Lascarides, 2008) - Exemple artificiel)

Les auteurs n'ont pas observé cet effet pour les relations de type *result* et *explanation*. Cependant la suppression du connecteur *puisque* dans l'exemple 8 semble bien modifier la relation inférée. Le connecteur marquait le fait que la seconde proposition est une explication pour ce qui est énoncé dans la première ce qui implique que l'événement de migration intervient avant le fait de devenir des adversaires. Sans le connecteur, l'ordre des événements semble inversé et suit l'ordre du texte. De plus une relation de *résultat* paraît inférable, le premier segment expliquant le second. Ici ce n'est donc pas une relation implicite déjà présente que l'on infère en supprimant le connecteur mais une nouvelle relation. On peut peut-être expliquer cette modification par l'hypothèse d'inférence d'une continuité linéaire par défaut liée à un manque de connaissance dans le domaine, continuité liée à une interprétation causale car le contenu des arguments le permet. On note que le localisateur temporel « alors » est modifié par la suppression du connecteur : avec le connecteur il semble porter sur toute la phrase, sans, il porte plutôt sur la seconde proposition.

(8) { [Les Amorrites deviennent à la période suivante de sérieux adversaires des souverains d'Ur,] [(***puisqu'***) ils commencent alors à migrer en grand nombre vers la Mésopotamie.] }*explanation*

(ANNODIS - Document WK_-_amorrites)

3.1.4. *Effets sur un système de classification*

On aimerait pouvoir définir des critères permettant de choisir les exemples artificiels les plus susceptibles d'aider à l'identification des exemples de relations implicites, par exemple en termes de connecteurs ou de relations. D'après les considérations précédentes, on pourrait penser que nous ne devrions pas prendre en compte les contrastes de type violation d'attente, puisqu'ils ne sont jamais implicites. Cependant les indices qu'ils mettent en jeu, comme des oppositions lexicales (*love*/*hate*) peuvent apparaître dans les exemples de contraste formel. Les cas plus gênants pour un système de classification sont ceux où la relation inférable est modifiée, puisque nous pourrions avoir des exemples de relations implicites naturels et artificiels similaires mais annotés avec des relations différentes. Cependant, la modification de la relation dépend fortement du contenu des segments liés, critère difficile à encoder. Comme ces cas sont rares (1 cas sur 28 exemples de *explication* considérés), nous espérons que la stratégie de sélection d'exemples ainsi que les méthodes de combinaison permettront de gérer ce problème.

3.2. *Différence de distribution*

En apprentissage statistique, on fait en général l'hypothèse que données d'entraînement et de test sont identiquement et indépendamment distribuées (*i.i.d.*). Ce qui signifie, informellement, que les données suivent une même loi de probabilité (généralement inconnue) et que leur échantillonnage a été fait de manière aléatoire.

On voit clairement, de par sa conception, que la méthode de création automatique d'exemples implicites proposée par Marcu et Echihabi (2002) risque fort de mettre à mal cette hypothèse importante. Rien en effet dans cette méthode ne garantit qu'on obtienne, pour nos données artificielles, une distribution proche de celle des données naturelles. Ce qui pose en retour le problème d'un apprentissage dans un contexte où les données ne sont pas *i.i.d.* On a ainsi deux ensembles de données qui se ressemblent (même ensemble d'étiquettes, les instances sont des segments de texte) mais qui sont néanmoins distribués différemment, et ce, à plusieurs titres. D'une part, les données artificielles sont par définition obtenues à partir d'exemples de relations explicites : il n'y a aucune garantie que ces données soient distribuées comme les exemples de relations implicites attestés. La différence porte tant sur la distribution des étiquettes (les relations) que sur l'association entre étiquettes et entrées (les paires des segments) à classer. En outre, la suppression du connecteur modifie les exemples, ce qui peut avoir une incidence comme noté dans la section précédente. Enfin, l'annotation automatique est fondée sur des heuristiques permettant d'identifier les connecteurs et leurs arguments. Ces heuristiques induisent un bruit supplémentaire, que l'on ne retrouve pas dans les données naturelles. Par exemple, on peut s'être trompé sur l'identification d'un connecteur : la forme n'était pas en emploi discursif, on n'a en fait aucune relation ou alors une relation différente de celle annotée. On peut aussi avoir fait des erreurs au niveau de l'identification des arguments. Nous décrivons dans la suite de cette

section les différences de distribution en reprenant les catégories décrites en adaptation de domaine, par exemple dans (Moreno-Torres *et al.*, 2012) et (Jiang, 2008).

3.2.1. *Déséquilibre des classes (*class imbalance *ou* prior probability drift*)*

Dans ce cas, la différence porte sur la distribution marginale des classes. Déjà différente entre données explicites et implicites naturelles, notre heuristique en produit une nouvelle. Les chiffres exacts sont donnés dans la section 5. La classe sousreprésentée dans les données naturelles, *contraste*, devient surreprésentée dans les données artificielles. Pour cette classe, la forme *mais*, toujours en emploi discursif, a permis d'extraire 75 % des données. En revanche, la relation *continuation* devient sousreprésentée dans les données artificielles. Les connecteurs de cette relation, comme *et*, sont plus ambigus en emploi et nous avons dû définir des motifs plus stricts pour éviter de récupérer de mauvais exemples. Cette différence peut être facilement gérée en réechantillonnant les données artificielles suivant la distribution des données implicites naturelles.

3.2.2. *Distribution des paires de segments (*covariate shift *ou* population drift*)*

Ici la différence porte sur la distribution marginale des *inputs*, les paires de segments en entrée. Le fait d'utiliser des exemples explicites induit une différence, car on peut penser que sans connecteur les indices utilisés sont différents. De plus, la suppression du connecteur peut aboutir à des exemples agrammaticaux. La segmentation induit aussi des différences. D'une part, on a potentiellement des erreurs de segmentation dans les données artificielles. D'autre part, la segmentation des données artificielles correspond à des hypothèses simplificatrices : un argument couvre au plus une phrase et on a au plus deux arguments par phrase. La segmentation des données naturelles ne suit bien sûr pas ces hypothèses : les arguments peuvent être aussi multiphrastiques ou séparer une phrase en plus de deux segments. Enfin, on a potentiellement un biais en termes de genre : les exemples artificiels sont tous construits à partir de l'*Est Républicain* mais les données naturelles proviennent aussi de Wikipédia (voir la description des données en section 5).

3.2.3. *Association entre étiquettes et paires de segments (*concept drift *ou* functional relation change*)*

Dans ce cas la différence porte sur la distribution jointe entre classes et paires de segments. On peut se rendre compte de la différence de distribution sur l'association entre étiquettes et exemples en considérant certaines caractéristiques des données. On peut par exemple regarder la répartition entre occurrences de relations inter et intraphrastiques (la relation s'établit entre deux phrases ou deux segments à l'intérieur d'une phrase), voir tableau 1. Entre occurrences de relations implicites naturelles et artificielles, on a une proportion d'interphrastiques similaire pour *contraste* (57,1 % d'interphrastiques dans les deux types de données), proche pour *résultat* (45,7 % d'interphrastiques dans les données naturelles, 39,8 % dans les artificielles) mais très différente pour *continuation* (70,0 % d'interphrastiques dans les naturelles, 96,5 % dans

les artificielles), et pour *explication* (21,4 % dans les naturelles, 53,0 % dans les artificielles). On observe aussi que la proportion d'occurrences de relations interphrastiques dans les données artificielles ne reflète pas celle des données explicites, ceci étant dû à notre heuristique.

Relations	Implicites	Explicites	Artificiels
contraste	57,1 % (57,1 %)	40,0 %	57,1 %
résultat	50,9 % (45,7 %)	65,4 %	39,8 %
continuation	66,9 % (70,0 %)	52,5 %	96,5 %
explication	21,4 % (21,4 %)	37,9 %	53,0 %
Total	494 (252)	614	392 260

Tableau 1. *Répartition des occurrences de relations interphrastiques implicites (naturelles), explicites et artificielles pour tous les exemples disponibles, (X %) pour les seuls exemples utilisés dans nos expériences dans le cas des implicites*

La stratégie d'annotation automatique peut aboutir à des erreurs d'étiquetage. On a pu identifier une forme comme connecteur alors qu'elle n'était pas en emploi discursif : soit il n'y a en fait aucune relation entre les arguments, soit une autre relation était présente. De plus la suppression du connecteur peut avoir pour effet de modifier la relation identifiable comme vu dans la section précédente. Dans ces cas, il est possible que nos données implicites naturelles contiennent des exemples ressemblant à ces exemples artificiels mais annotés avec la relation que l'on identifie sans prendre en compte le connecteur.

4. Modèles

Dans des études précédentes, l'entraînement sur les seules données artificielles aboutit à des résultats inférieurs à un entraînement sur des données naturelles (pourtant bien moins nombreuses). Ceci s'explique par les différences de distribution entre les deux ensembles de données décrites dans la section précédente. Dans cette section, nous décrivons différentes méthodes visant à exploiter les nouvelles données artificielles, non plus seules, mais en combinaison avec les données naturelles existantes.

4.1. *Cadre classique de classification supervisée*

La tâche de classification correspond à l'assignation à un exemple d'une classe parmi un ensemble prédéfini. En apprentissage supervisé, on apprend la fonction de classification à partir d'un ensemble de données associées à leur classe, les données d'entraînement. Dans nos expériences, nous utilisons un modèle par maximum d'entropie, ou régression logistique (Berger *et al.*, 1996), qui donne de bonnes performances pour différents problèmes de TAL. C'est un modèle probabiliste, on obtient une distribution de probabilité sur les classes pour chaque exemple, et un modèle

discriminant, on modélise directement la probabilité conditionnelle des classes étant donnée les données. Ce type de modèle a aussi l'avantage de permettre l'ajout de nombreux descripteurs potentiellement redondants sans faire d'hypothèse d'indépendance.

Pour décrire formellement ce modèle, on note $\mathcal{X}$ l'ensemble des données en entrée et $\mathcal{Y}$ les sorties possibles. On dispose d'un ensemble de données étiquetées $S = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$. On veut apprendre une fonction hypothèse $h : \mathcal{X} \rightarrow \mathcal{Y}$. La prédiction se fonde sur une représentation des données sous forme de vecteurs de traits, $\Phi(x_i, y_i) \in \mathbb{R}^d$ (d le nombre de dimensions ou traits), et sur un vecteur de poids $\mathbf{w} \in \mathbb{R}^{\mathbf{d}}$, les paramètres du modèle. À chaque trait ϕ_i est associé un poids w_i. Le vecteur de paramètres est appris à l'entraînement à partir d'un ensemble de données étiquetées $S_{train} \subset S$. Dans un modèle par maximum d'entropie, la classe est prédite selon la formule [1], pour une instance $x \in \mathcal{X}$.

$$h(x) = \text{argmax}_{y \in \mathcal{Y}} \frac{\exp(\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}, \mathbf{y}))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}, \mathbf{y}'))} \quad [1]$$

L'apprentissage doit permettre de déterminer le « meilleur » vecteur de paramètres $\hat{\mathrm{w}}$, entendu ici comme celui qui maximise la log-vraisemblance des données d'entraînement et, afin d'éviter une trop grande attache aux données, on ajoute un terme de régularisation qui va pénaliser les poids trop grands. On aboutit à la formule [2], où λ est le paramètre de régularisation du modèle, plus précisément, c'est la précision de la *prior* gaussienne sur les poids.

$$\hat{\mathrm{w}} = \text{argmax}_w \sum_{i=1}^{n} \log(P_w(y_i|x_i)) - \frac{\lambda}{2} \sum_{j=1}^{n} w_j^2 \quad [2]$$

Différents algorithmes peuvent être utilisés pour estimer ces paramètres, on utilise ici une méthode d'optimisation dite à mémoire limitéee (*limited memory BFGS*) implémentée dans MegaM[7].

4.2. *Cadre de l'adaptation de domaine*

Le problème de différence de distribution entre les deux types de données nous a amenés à étudier les méthodes utilisées en adaptation de domaine. En effet, l'adaptation de domaine fait réfèrence à des situations d'apprentissage où l'on dispose d'ensembles de données suivant des distributions différentes. Par exemple, on dispose de données annotées en catégories morphosyntaxiques dans le domaine journalistique et on veut annoter des données issues de manuels informatiques. Si l'on utilise simplement un modèle construit sur les premières données, on observe une dégradation des

7. `http://www.umiacs.umd.edu/~hal/megam/version0_3/`

performances : la différence en termes de domaine conduit à une baisse de performance, le modèle ne parvient pas à une bonne généralisation du premier domaine au second.

Plus formellement, on dispose d'un ensemble de données annotées dites sources D_S, en général en grande quantité, et d'un ensemble de données cibles D_T. On suppose en général qu'une petite partie des données cibles est aussi annotée (mais il existe aussi des techniques qui ne font pas cette hypothèse, voir (Daumé III *et al.*, 2010)). Les données sources suivent une distribution P_S sous-jacente inconnue, et les données cibles une distribution sous-jacente inconnue P_T. Les méthodes développées dans le cadre de l'adaptation de domaine se sont intéressées à différents cas, selon les différences distributionnelles mises en jeu présentées à la section précédente. Le problème de déséquilibre des classes est bien défini et géré par des méthodes de rééchantillonnage (*up-* ou *down-sampling*) ou de pondération des exemples. Les deux autres cas, *covariate shift* et *change of functionnal relation* sont moins clairement séparés (Sogaard, 2013).

Les problèmes d'adaptation de domaine ont été traités dans de nombreuses études et diverses solutions ont été proposées. Daumé III (2007) décrit un système qui se fonde sur la représentation des données, la divisant en trois espaces, un pour chaque domaine et un commun correspondant aux caractéristiques partagées. Ainsi sur la tâche de l'étiquetage morphosyntaxique, *the* aura la catégorie déterminant quel que soit le domaine tandis que *monitor* sera plutôt un verbe dans les articles de journaux et plutôt un nom dans des manuels informatiques. Il compare cette stratégie à un ensemble de systèmes dits de *baseline*, certains étant jugés difficiles à battre. Ce sont ces modèles que nous testons et que nous décrivons ci-dessous. Une autre approche a été proposée par Chelba et Acero (2006) : ces auteurs construisent un modèle sur les données sources puis l'utilisent comme régularisation pour un modèle construit sur les données cibles. On donne ainsi au modèle une information *a priori* sur les données et on gère une contrainte de proximité entre les deux modèles à travers un paramètre de régularisation. D'autres méthodes ont été proposées, notamment Jiang (2008) décrit des stratégies reposant soit sur une estimation de la différence entre les probabilités marginales des données soit sur une forme de filtrage des exemples sources qui induisent le modèle en erreur, ceux qui sont trop différents en termes de probabilité conditionnelle.

Nous ne nous plaçons pas exactement dans un cadre d'adaptation de domaine, au sens où nous ne nous intéressons pas particulièrement à la différence en termes de genre entre nos données (Wikipédia ou articles journalistiques), mais surtout parce que nous ne cherchons pas à construire un modèle performant sur les données artificielles et sur les données naturelles, seules les secondes nous intéressent. Cependant, ce cadre correspond à une situation d'apprentissage à partir de données tirées d'une distribution différente de celle des données d'évaluation. Dans ce sens, les données annotées automatiquement constituent nos données sources, et les données manuellement annotées, sur lesquelles on veut améliorer les performances, nos données cibles.

4.3. *Modèles testés*

Nous avons commencé par les méthodes les plus simples proposées par Daumé III (2007), des méthodes décrites comme relativement compétitives et simples à mettre en œuvre. Les différentes méthodes de combinaison que nous proposons diffèrent selon que la combinaison s'opère directement au niveau des jeux de données ou au niveau des modèles entraînés sur ceux-ci. Nous avons ajouté un principe de sélection automatique des exemples automatiquement annotés fondé sur la confiance du modèle entraîné sur ces données concernant sa prédiction afin de gérer une partie du bruit, et donc de sélectionner les exemples les plus sûrs. Les performances de tous les systèmes seront comparées à celles des systèmes de base entraînés séparément sur les deux ensembles de données dans la section 6.

4.3.1. *Systèmes de base*

Le premier système de base (NATONLY) est celui où l'on s'entraîne et où l'on s'évalue sur les données naturelles. Ce système nous donne les performances à dépasser et correspond à un cadre classique d'apprentissage où toutes les données sont supposées être tirées identiquement de la même distribution sous-jacente inconnue.

Le second système de base (ARTONLY) est celui où l'on s'entraîne sur les données annotées automatiquement (ou artificielles) et où l'on s'évalue sur les données naturelles. C'est le système proposé dans les études de Blair-Goldensohn *et al.* (2007) et de Sporleder et Lascarides (2008) et nous devrions normalement observer, comme dans ces études, une baisse importante des performances.

Ces systèmes de base servent de comparaison aux autres systèmes testés, systèmes où sont combinés les données ou les modèles.

4.3.2. *Combinaison des données*

La première stratégie de combinaison que nous étudions (UNION) relève du premier type : elle consiste à créer un corpus d'entraînement qui contient la réunion des deux ensembles de données. Cette méthode ne permet pas de gérer l'importance de chacun des deux ensembles de données ou d'en estimer l'impact sur le système. Nous la déclinons donc en deux autres stratégies que nous décrivons *infra*.

Une première stratégie dérivée (ARTSUB) consiste à prendre, non pas l'intégralité des données artificielles, mais des sous-ensembles aléatoires de ces données, en addition des données naturelles. Cette méthode est un peu plus subtile dans la mesure où l'on peut faire varier la proportion des exemples artificiels par rapport aux exemples naturels. On ne prend cependant pas en compte l'intégralité des données artificielles donc on perd de l'information.

Enfin, la troisième méthode du premier type, dérivée de la première, (NATW) garde cette fois la totalité des données artificielles mais pondère (ou duplique) les exemples naturels de manière à éviter un déséquilibre trop grand au profit des données artificielles.

4.3.3. *Combinaison des modèles*

Dans le second type de méthode, nous avons tout d'abord une méthode (ADDPRED) qui consiste à utiliser les prédictions d'un modèle entraîné sur les données artificielles (à savoir les données « sources ») comme descripteur dans le modèle entraîné sur les données naturelles (à savoir les données « cibles »). Le paramètre associé à ce descripteur mesure donc l'importance à accorder aux prédictions du modèle entraîné sur les données artificielles. Cette méthode est la meilleure *baseline* et le troisième meilleur modèle dans (Daumé III et Marcu, 2006).

Une variation de cette méthode (ADDPROB) utilise en plus le score de confiance (ici la probabilité) du modèle artificiel comme descripteur supplémentaire dans le modèle construit sur les données naturelles. L'idée à la base de ces méthodes est que même si le classifieur se trompe, il est peut-être consistant dans ses erreurs qui peuvent se révéler une source d'information utile pour le modèle.

Une troisième méthode (ARTINIT) vise à initialiser les paramètres du modèle entraîné sur les données naturelles avec ceux du modèle utilisant les données artificielles. Cette méthode permet de fournir une information *a priori* au modèle entraîné sur les données naturelles, plutôt que de l'initialiser aléatoirement. Enfin, la dernière méthode (LININT) se fonde sur une interpolation linéaire de deux modèles préalablement entraînés sur chacun des ensembles de données. Les paramètres de l'interpolation linéaire permettent de faire varier l'importance de chacun des modèles.

4.3.4. *Sélection automatique d'exemples*

Nous avons aussi testé toutes les stratégies en ajoutant une étape de sélection automatique d'exemples artificiels. La méthode utilisée est naïve puisqu'elle se fonde simplement sur la probabilité de l'étiquette prédite : nous testons différents seuils sur ces probabilités en ajoutant à chaque fois les seuls exemples prédits avec une probabilité supérieure au seuil, et en rééquilibrant l'ensemble des données. Cette sélection vise à écarter des données bruitées, en explorant finalement l'une des voies proposées par Marcu et Echihabi (2002) et développée d'une autre manière par Blair-Goldensohn *et al.* (2007), à savoir améliorer la qualité du corpus artificiel. Cette méthode permet aussi dans une certaine mesure de sélectionner les données pour lesquelles l'hypothèse de redondance du connecteur est vérifiée : plus le modèle est confiant dans sa prédiction, plus on a de chances qu'il ait trouvé de bons indices pour sa prédiction.

4.4. *Jeu de traits*

Notre jeu de traits se fonde sur les travaux existants avec quelques adaptations notables pour le français. Ces traits exploitent des informations de surface, ainsi que d'autres issues d'un traitement linguistique plus profond. Par comparaison, Marcu et Echihabi (2002) ne se fondent que sur la cooccurrence de lemmes dans les segments. Sporleder et Lascarides (2007) montrent que la prise en compte de différents types de traits linguistiquement motivés améliore les performances. Sporleder et Lascarides

(2008) utilisent des traits variés dont des bigrammes de lemmes mais sans traits syntaxiques. Nous avons testé des traits lexicosyntaxiques utilisés dans les précédentes études sur cette tâche. Nous n'avons pas pu reprendre les traits sémantiques comme les classes sémantiques des têtes des arguments (déterminées par l'analyse en dépendance) car les ressources nécessaires n'existent pas pour le français.

Les traits dont les valeurs sont discrètes ou catégorielles ont été binarisés.

Certains traits sont calculés pour chaque argument :

– **indice de complexité syntaxique** : nombre de syntagmes nominaux, verbaux, prépositionnels, adjectivaux, adverbiaux (continu) ;

– **information sur la tête d'un argument** :

- lemme d'éléments négatifs dépendant de la tête, du type « pas, aucun... » (booléen),

- informations temporelles/aspectuelles : nombre de fois où un lemme de fonction auxiliaire dépendant de la tête apparaît (continu), temps, personne, nombre de l'auxiliaire (booléen),

- informations sur les dépendants de la tête : présence d'un dépendant de la tête, du sujet ou de l'objet selon son type dans l'analyse en dépendance (objet, « par »objet, modifieur ou dépendant prépositionnel) (booléen) ; catégorie morphosyntaxique des modifieurs et des dépendants prépositionnels de la tête, du sujet ou de l'objet (booléen),

- informations morphologiques : temps et personne de la tête verbale, genre de la tête non verbale, nombre de la tête, catégorie morphosyntaxique précise (par exemple « VPP ») et simplifiée (respectivement « V ») (booléen).

D'autres traits portent sur la paire d'arguments :

– **trait de position** : si la relation s'établit entre deux phrases (interphrastique) ou entre deux propositions à l'intérieur d'une phrase (intraphrastique) (booléen) ;

– **indice de continuité thématique** : chevauchement en lemmes et en lemmes de catégorie ouverte (continu) ;

– **information sur les têtes des arguments** :

- paire formée des temps verbaux des têtes (booléen),

- paire formée des nombres des têtes (booléen).

Notons finalement que notre but portant avant tout sur la combinaison de données, nous n'avons pas cherché à optimiser ce jeu de traits, ce qui aurait introduit un paramètre supplémentaire dans notre modèle.

5. Données

5.1. *Choix des relations*

Nous avons choisi de nous restreindre à 4 relations : *contraste*, *résultat*, *continuation* et *explication*. Ces relations sont annotées dans le corpus français utilisé et correspondent à des exemples de relations implicites et explicites. De plus ce sont 4 des 5 relations (*summary/résumé* n'est pas annotée dans ANNODIS) utilisées dans (Sporleder et Lascarides, 2008), ce qui nous permet une comparaison, cependant non directe puisque la langue est différente. Enfin, ce sont des relations qui peuvent s'exprimer de manière implicite (contrairement, par exemple, à la relation de type *condition* qui nécessite une lexicalisation par un connecteur), mais aussi de manière explicite, ce qui est nécessaire pour construire le corpus artificiel (contrairement, par exemple, à la relation *encadrement* pour laquelle nous ne disposons pas d'un lexique de connecteurs).

Dans nos données naturelles, nous avons fusionné les métarelations (ou relations pragmatiques, qui s'établissent entre des actes de parole) avec les relations correspondantes (non pragmatiques, qui s'établissent entre des faits) avec l'hypothèse qu'elles mettaient en jeu le même genre d'indices et de constructions. Les données naturelles permettent d'obtenir des exemples de relations implicites manuellement annotés. Les données générées automatiquement sont des exemples explicites extraits par heuristique de données brutes dans lesquels on supprime le connecteur, on a alors des données implicites artificielles.

5.2. *Données annotées manuellement : le corpus* ANNODIS

Le projet ANNODIS (Afantenos *et al.*, 2012) vise la construction d'un corpus annoté en discours pour le français à deux niveaux : d'une part sont annotées des relations de discours entre EDU et segments complexes suivant le cadre SDRT (microstructure), d'autre part, sont annotées des structures discursives de haut niveau concernant l'organisation fonctionnelle du document (macrostructure). Ici, nous nous intéressons uniquement au niveau microstructurel. La version du corpus utilisée (en date du 15/11/2012) comporte 86 documents provenant de l'*Est Républicain* et de Wikipédia. Au total, 3 339 exemples sont annotés avec 17 relations rhétoriques. On note que le manque de données manuellement annotées est encore plus important pour le français que pour l'anglais, puisqu'en ne prenant en compte que le corpus anglais du PDTB (environ 40 000 exemples), on dispose d'environ 12 fois moins d'exemples annotés manuellement. Les documents sont segmentés en EDU : propositions, syntagmes prépositionnels, adverbiaux détachés à gauche et incises, si le segment contient la description d'une éventualité. Les relations sont annotées entre EDU ou segments complexes, contigus ou non. Les connecteurs discursifs ne sont pas annotés.

Le corpus a subi une série de prétraitements. Le MElt tagger (Denis et Sagot, 2009) fournit un étiquetage en catégories morphosyntaxiques, une lemmatisation, des indi-

cations morphologiques (temps, personne, genre, nombre). Le MSTParser (Candito *et al.*, 2010) fournit une analyse en dépendances. Afin de restreindre notre étude aux relations implicites, nous utilisons le *LexConn*, lexique des connecteurs discursifs du français développé par Roze *et al.* (2012) et étendu en 2012 aux connecteurs introduisant des syntagmes nominaux. Nous utilisons une méthode simple : nous projetons le lexique (sauf la forme *à* jugée trop ambiguë) sur les données, ce qui nous permet d'identifier tout token correspondant à un connecteur. Nous ne contraignons pas cette identification sur des critères de position. Cette méthode nous assure d'identifier tout connecteur donc de ne récupérer que des exemples de relations implicites mais comporte le risque d'en perdre certains. Sur les 1 108 exemples disponibles pour les 4 relations, nous disposons de 494 exemples de relations implicites ; la distribution des exemples par relation est résumée dans le tableau 2.

Relations	Explicites	Implicites	Total
contraste	100	42	142
résultat	52	110	162
continuation	404	272	676
explication	58	70	128
Total	614	494	1 108

Tableau 2. *Corpus* ANNODIS *: nombre d'exemples de relations explicites et implicites par relation*

5.3. *Données annotées automatiquement*

Nous utilisons le lexique de connecteurs *LexConn* pour extraire automatiquement des exemples de l'une des 4 relations discursives considérées dans le corpus composé d'articles de l'*Est Républicain* (9 M de phrases), avec les mêmes traitements que pour ANNODIS.

Le lexique *LexConn* contient en tout 329 formes, dont 131 ne peuvent exprimer qu'une seule relation parmi les 4 que nous avons choisies. Les connecteurs ambigus en termes de la relation déclenchée pourraient aussi être intéressants mais ils amèneraient du bruit dans l'annotation automatique puisque nous ne disposons pas de modèle pour les désambiguïser. Nous avons fusionné les relations et les métarelations correspondantes, avec l'hypothèse qu'elles mettent en jeu les mêmes indices. Nous avons aussi regroupé les 3 relations de type contrastif définies dans ce lexique comme cela a été fait dans le corpus manuellement annoté. Nous n'avons pas pris en compte 3 connecteurs qui ne sont associés à aucune catégorie morphosyntaxique dans le lexique (catégorie morphosyntaxique inconnue, comme *d'où que*). Après une première évaluation, nous avons choisi de supprimer 6 connecteurs particulièrement ambigus en emploi pour lesquels il était difficile de définir des motifs départageant emploi discursif et non discursif (comme *maintenant*). Nous disposons finalement de 122 connecteurs, seuls 100 ont effectivement été utiles finalement, les 22 non utili-

sés soit n'ont pas été trouvés dans le corpus soit n'ont pas été trouvés pas dans une configuration définie comme discursive.

L'heuristique d'annotation automatique des exemples se fait en deux étapes. Nous identifions d'abord les formes en emploi discursif en utilisant des motifs (voir tableau 3) définis manuellement pour chaque forme de connecteur. Ces motifs sont fondés sur la position de la forme (initiale, médiane ou finale), sa catégorie morphosyntaxique (POS) et la ponctuation autour de la forme, et en nous aidant des indications de *LexConn*. Nous identifions ensuite les arguments d'un connecteur, ce qui peut être vu comme une simplification du problème de segmentation. Nous faisons les mêmes hypothèses simplificatrices que dans les études précédentes : les arguments sont adjacents et couvrent au plus une phrase ou au plus deux EDU par phrase. Nous utilisons aussi la position du connecteur, sa catégorie morphosyntaxique et la ponctuation pour cette identification.

Positions	POS	Motifs	Exemples
Interphrastique	Toutes POS	A1. **C**(,) A2.	A1. **Malheureusement**(,) A2 A1. **Surtout**, A2.
	Adv.	A1. beg-A2(,) **C**(,) end-A2. A1. A2, **C**.	A1. beg-A2, **de plus**, end-A2. A1. beg-A2(,) **en outre**(,) end-A2. A1. A2, **remarque**.
Intraphrastique	Toutes POS	A1, **C**(,) A2.	A1, **de plus**(,) A2. A1(,) **donc**(,) A2.
	CS et Prep.	**C** A1, A2.	**Preuve que** A1, A2. **Puisque** A1, A2.
	Adv.	A1, beg-A2(,) **C** (,) end-A2. A1, A2, **C**.	A1, beg-A2, **de plus**, end-A2. A1, beg-A2(,) **en outre**(,) A2. A1, A2, **réflexion faite**.

Tableau 3. *Les motifs définis et quelques exemples correspondants. « A1 » correspond au premier argument, « A2 » au second et « C » au connecteur ; « beg » et « end » correspondent respectivement au début et à la fin d'un argument ; « (x) » indique le caractère facultatif de « x », selon la forme des connecteurs mis en jeu. Certains motifs ne sont applicables que pour certaines catégories morphosyntaxiques de connecteur (« POS ») entre conjonction de subordination (CS), préposition (Prep.) et adverbe (Adv.).*

Cette méthode simple permet de générer rapidement de gros volumes de données : au total, nous avons pu extraire 392 260 exemples (voir tableau 4). Lorsque deux connecteurs sont présents dans un segment, il peut arriver que l'un modifie l'autre ou que l'on ait en fait un double connecteur (par exemple « *mais* **parce qu'**il est... »). Dans ce cas, nous risquons de récupérer les mêmes arguments pour deux formes déclenchant des relations différentes ce qui est problématique pour un système de classification. Quand 2 connecteurs sont identifiés dans 1 phrase, nous générons 2 exemples à condition que les arguments soient différents : 1 exemple doit correspondre à 1 relation s'établissant entre 2 phrases et l'autre à 1 relation s'établissant entre 2 propositions à l'intérieur d'1 phrase. Nous avons équilibré le corpus en relations en conservant le

maximum d'exemples disponibles en un corpus d'entraînement (80 % des données), un de développement (10 %) et un de test (10 %).

Notons quelques différences importantes de distribution entre les données naturelles et artificielles : *continuation* la plus représentée dans les naturelles devient la moins représentée dans les artificielles. Ceci est dû à la forte ambiguïté des connecteurs de cette relation qui nous ont forcés à définir des motifs stricts pour l'extraction des exemples. Notons finalement que cette méthode génère du bruit : sur 250 exemples choisis aléatoirement, on trouve 37 erreurs de frontière d'arguments et 18 d'emplois non discursifs.

`Relations`	Disponibles	Entraînement	Développement	Test	Total
contraste	252 793	23 409	2 926	2 926	29 261
résultat	50 297	23 409	2 926	2 926	29 261
continuation	29 261	23 409	2 926	2 926	29 261
explication	59 909	23 409	2 926	2 926	29 261
`Total`	392 260	93 636	11 704	11 704	117 044

Tableau 4. *Corpus artificiel : nombre d'exemples par relation*

6. Expériences

Pour rappel, l'objectif central de ces expériences est de déterminer dans quelle mesure l'ajout de données artificielles, *via* les différentes méthodes présentées en section 4.3, peut nous permettre de dépasser les performances obtenues en s'entraînant sur des données naturelles présentes seulement en faible quantité.

Les expériences sont réalisées avec l'implémentation de l'algorithme par maximum d'entropie fournie dans la librairie MegaM en version multiclasse avec au maximum 100 itérations. Nous n'avons pas cherché à optimiser le coefficient de régularisation, sa valeur est de 1.

Nous utilisons un corpus de données naturelles équilibré à 70 exemples au maximum par relation. Il faudra envisager des expériences conservant la distribution naturelle des données, très déséquilibrée, mais pour l'instant nous nous focalisons sur l'aspect combinaison des données. Étant donné le faible volume de données manuellement annotées disponibles, il n'est pas possible de découper celles-ci en ensembles distincts pour l'entraînement et la validation. En conséquence, nous utilisons une validation croisée dite « enchâssée » (*nested cross-validation*). Cette méthodologie permet de s'assurer que l'optimisation des hyperparamètres (décrits en section 6.2) se fait sur des données différentes de celles sur lesquelles on s'évalue. Il a été montré que cette méthode permet d'obtenir une estimation réaliste de l'erreur (Scheffer, 1999 ; Varma et Simon, 2006). Comme son nom l'indique, la validation croisée enchâssée comporte

2 boucles : une boucle interne de validation croisée qui sert à optimiser le ou les hyperparamètres, donc à sélectionner le modèle, et une boucle externe de validation croisée qui sert à mesurer les performances obtenues, donc à estimer son erreur.

La boucle externe parcourt les N sous-ensembles du premier découpage aléatoire. Le sous-ensemble k, avec $1 \leq k \leq N$, est considéré comme l'ensemble d'évaluation et les $N - 1$ autres sous-ensembles servent à la sélection du modèle dans la boucle interne : l'ensemble des données correspondant à ces $N - 1$ sous-ensembles est découpé aléatoirement en M sous-ensembles et une validation croisée sur ces M sous-ensembles permet de déterminer les meilleures valeurs pour les hyperparamètres. Un classifieur est enfin entraîné avec les $N - 1$ sous-ensembles de la boucle externe et évalué sur le sous-ensemble k en utilisant les meilleures valeurs pour les hyperparamètres. Nous avons utilisé ici 2 validations croisées en 5 sous-ensembles pour déterminer et évaluer les meilleurs modèles pour chacun des systèmes décrits en section 4.3. Il n'est donc pas garanti que les meilleurs soient sélectionnés à chaque étape de test, mais cette méthode permet d'évaluer la stabilité du système par rapport au choix des hyperparamètres (*i.e.*, les valeurs ne doivent pas être trop éparpillées), le surentraînement (*i.e.*, si les estimations dans les boucles internes et externes sont très éloignées, c'est que la méthode d'optimisation conduit à un surentraînement) et la stabilité des modèles (*i.e.*, on s'intéresse à la variance dans la capacité prédictive, sur les résultats sur les sous-ensembles de la boucle externe). Les systèmes n'utilisant aucun hyperparamètre sont simplement évalués dans la boucle externe.

Comme dans les études précédentes, les performances sont données en termes d'exactitude globale (micromoyenne) sur l'ensemble des relations, des scores ventilés de mesure F1 par relation sont également fournis. Dans le cas de très petits échantillons (ici $N = 5$), il est difficile de calculer la significativité des écarts de performance observés. (de Winter, 2013) montre que le test de Student appairé peut être utilisé dans ce cas, à condition que l'effet de taille (*effect size*, calculé avec le coefficient de Cohen) et que la corrélation entre les échantillons soient suffisamment importants. (de Winter, 2013) indique qu'au contraire, le test des rangs signés de Wilcoxon peut conduire à des p-valeurs trop hautes avec de si petits échantillons. Nous choisissons donc de donner les résultats en utilisant le test de Student (avec une p-valeur $< 0{,}05$).

6.1. *Modèles de base*

Dans un premier temps, nous construisons deux modèles distincts, l'un à partir des seules données naturelles (NATONLY, 252 exemples), l'autre à partir des seules données artificielles (ARTONLY, 93 636 exemples d'entraînement). Notre modèle NATONLY obtient une exactitude de 37,3 %, avec des scores de F1 par relation compris entre 15,0 % pour *contraste* et 47,9 % pour *explication* (voir tableau 5). La relation *contraste* est donc très mal identifiée peut-être parce que sous-représentée, seulement 42 exemples contre 70 pour les autres relations, le manque de données joue probablement ici un rôle important.

Le modèle ARTONLY obtient une exactitude de 47,8 % lorsque évalué sur le même type de données (11 704 exemples de test), mais de 23,0 % lorsque évalué sur les données naturelles (voir tableau 5). Cette baisse importante est comparable à celle observée dans les études précédentes sur l'anglais. Elle s'explique par les différences de distribution étudiées en section 3.2. De manière générale, nous observons des dégradations par rapport à NATONLY pour l'identification de *résultat* et *explication* (voir tableau 5). En revanche l'identification de *contraste* présente une amélioration, obtenant 23,2 % de F1 avec 11 exemples correctement identifiés contre 6 précédemment.

	NATONLY	ARTONLY	
Données de test	Naturelles	Naturelles	Artificielles
Exactitude	37,3	23,0	47,8
contraste	15,0	23,2	38,3
résultat	47,6	15,7	57,4
continuation	28,1	32,1	54,3
explication	47,9	22,4	37,5

Tableau 5. *Modèles de base, exactitude du système et F1 par relation*

6.2. *Modèles avec combinaisons de données*

Dans cette section, nous présentons les résultats des systèmes qui exploitent à la fois les données naturelles et les données artificielles. Ces ensembles de données sont ou bien combinés directement ou bien donnent lieu à des modèles séparés qui sont combinés plus tard.

Certains de ces modèles utilisent des hyperparamètres. Ainsi, pour la pondération des exemples naturels nous testons différents coefficients de pondération c avec $c \in [0{,}5, 1, 5]$ et $c \in [10\,;2\,000]$ avec un incrément de 10 jusqu'à 100, de 50 jusqu'à 1 000 et de 500 jusqu'à 2 000. Pour l'ajout de sous-ensembles des données artificielles, nous ajoutons à chaque fois n exemples parmi ces données où $n = k$ fois le nombre de données naturelles disponibles avec $k \in [0{,}1\,;600]$ avec un incrément de 0,1 jusqu'à 1, de 10 jusqu'à 100 et de 50 jusqu'à 600. Enfin, pour l'interpolation linéaire des modèles, nous construisons un nouveau modèle en pondérant le modèle artificiel avec $\alpha \in [0{,}1\,;0{,}9]$ avec des incréments de 0,1 (dans le nouveau modèle, nous pondérons donc le modèle construit sur les données naturelles par un coefficient de $1 - \alpha$). Les scores des systèmes sont repris dans le tableau 6.

	UNION	NATW	ARTSUB	ADDPRED	ADDPROB	ARTINIT	LININT
Exactitude	22,6	38,9	34,5	39,3	38,9	40,1	39.3
contraste	22,0	20,3	15,0	16,0	15,6	16,9	17,1
résultat	15,2	40,4	39,4	50,6	48,0	45,9	45,4
continuation	38,2	44,7	36,2	31,9	31,9	34,0	38,0
explication	15,7	42,2	39,2	46,7	48,9	52,2	47,5

Tableau 6. *Modèles sans sélection d'exemples, exactitude du système et F1 par relation*

De manière générale, l'ensemble de ces systèmes avec les bons hyperparamètres conduit à des résultats au moins équivalents et parfois supérieurs en exactitude par rapport à NATONLY. Si la tendance générale est plutôt d'une hausse des performances, aucune des différences observées à ce stade ne semble cependant être statistiquement significative. Le meilleur score d'exactitude est obtenu avec le système ARTINIT (40,1 % d'exactitude, p-valeur de 0,18 avec un effet de taille faible, 0,39). Deux autres systèmes obtiennent un score d'exactitude supérieur à 39 %, ADDPRED et LININT, avec des résultats toujours non significatifs. Le modèle similaire à ADDPRED, mais exploitant en plus les probabilités (ADDPROB), mène à une légère diminution ce qui suggère que les traits de probabilité dégradent les performances. Pour ces systèmes, les scores d'exactitude obtenus sur chaque sous-ensemble sont très proches[8], notamment pour ADDPRED, ce qui suggère une bonne stabilité des modèles. Les autres systèmes nous permettent d'évaluer l'effet des données artificielles dans les résultats finaux.

La seule configuration qui mène à des résultats négatifs est l'union simple des corpus d'entraînement (UNION). Ce système obtient 22,6 % d'exactitude donc de l'ordre d'un entraînement sur les seules données artificielles. Ces résultats ne sont pas surprenants : les données naturelles étant environ 372 fois moins nombreuses que les données artificielles, elles se retrouvent pour ainsi dire « noyées » dans les données artificielles.

Les expériences de combinaison des données, ajout de sous-ensembles aléatoires des données artificielles (ARTSUB, 34,5 % d'exactitude) et pondération des exemples manuels (NATW, 38,9 % d'exactitude), montrent l'influence de l'importance relative entre les deux types de données sur l'exactitude des systèmes. Nous pouvons observer cet effet en regardant les scores obtenus dans la boucle interne d'optimisation : pour ces deux systèmes, les données naturelles doivent avoir un poids environ 2,5 fois supérieur aux données artificielles pour obtenir les meilleurs scores d'exactitude. La variance entre les valeurs d'hyperparamètres choisies est très haute pour ARTSUB, ceci est probablement dû au caractère aléatoire du choix des sous-ensembles qui peut conduire à des différences importantes. Elle est un peu moins forte pour NATW, montrant que cette méthode est plus robuste, mais le choix *a priori* d'une valeur pour

8. Écart-type de 0,074 pour ARTINIT pour une moyenne de 40,1 et de 0,037 pour ADDPRED et 0,061 pour ADDPROB pour une moyenne d'environ 39.

l'hyperparamètre reste large (1 020 à plus ou moins 272). L'interpolation linéaire (LINT) des deux modèles permet d'observer le même effet. Les meilleurs modèles sont obtenus avec un coefficient α en moyenne égal à $0{,}3$, donc de nouveau une influence plus de deux fois plus forte. L'écart-type est relativement élevé mais il semblerait raisonnable de choisir *a priori* pour un futur modèle une valeur correspondant à la moyenne.

Les méthodes de combinaison aboutissent à des systèmes d'exactitude similaire voire supérieure à NATONLY. Au niveau des scores par relation, nous avons observé lors de la phase d'optimisation, qu'une influence forte des données artificielles avait tendance à améliorer l'identification de *contrast* et de *continuation*. Ainsi la moyenne de la F1 augmente avec l'augmentation du coefficient α et l'interpolation linéaire des modèles. Au contraire, un poids fort des données artificielles entraîne une forte dégradation de l'identification de *résultat* et *explication*. La relation *contraste* profite peut-être de données artificielles moins bruitées : la majorité des exemples (plus de 75 %) sont construits à partir de *mais*, une forme qui est toujours en emploi discursif et dont les arguments sont dans l'ordre canonique, argument1 + connecteur + argument2. Les différences de performance au niveau des classes peuvent venir de distributions plus ou moins proches entre les deux types de donnée. En regardant la distribution en terme de traits (850 traits en tout), nous constatons un écart de plus de 30 % pour 2 et 5 traits pour *résultat* et *explication* mais aucun pour *contraste* et *continuation*, les relations pour lesquelles l'apport direct des données artificielles est positif.

6.3. *Modèles avec sélection automatique d'exemples*

Les expériences précédentes ont montré que l'ajout de données artificielles donnait le plus souvent lieu à des gains de performance, mais ces gains restent relativement modestes, voire non significatifs. Notre hypothèse est que de nombreux exemples artificiels amènent du bruit dans le modèle. Idéalement, nous souhaiterions être capables de sélectionner les exemples artificiels les plus informatifs et qui complémentent le mieux les données naturelles.

La méthode de sélection d'exemples que nous proposons a pour objectif d'éliminer les exemples potentiellement plus bruités. Pour cela, le modèle artificiel est utilisé sur les données d'entraînement et nous conservons les exemples prédits avec une probabilité supérieure à un seuil $s \in [0{,}3\,;0{,}85]$ avec un incrément de $0{,}1$ jusqu'à $0{,}5$ et de $0{,}05$ jusqu'à $0{,}85$. Si ce modèle est assez sûr de sa prédiction, on peut espérer que l'exemple ne correspond pas à du bruit, à une forme en emploi non discursif et/ou à une erreur de segmentation. Nous vérifions aussi, en quelque sorte, l'hypothèse de redondance du connecteur. Pour chaque seuil, nous rééquilibrons les données en nous fondant sur la relation la moins représentée (système + SELEC). Les scores des systèmes sont repris dans le tableau 7.

+ SELEC	UNION	NATW	ARTSUB	ADDPRED	ADDPROB	ARTINIT	LININT
Exactitude	40,1	41,3	39,3	**41,7***	35,7	36,9	36,5
contraste	25,9	19,2	21,6	20,8	14,5	18,9	17,5
résultat	45,3	48,3	44,4	51,0	41,1	34,5	38,2
continuation	34,8	32,4	33,8	31,2	30,2	31,0	37,8
explication	48,9	53,4	47,6	53,9	45,3	52,8	44,3

Tableau 7. *Modèles avec sélection d'exemples, exactitude du système et F1 par relation ; * signale un résultat significativement supérieur à* NATONLY

La sélection automatique d'exemples permet d'améliorer la plupart des résultats précédents, montrant l'intérêt de cette étape supplémentaire. En particulier, le système correspondant à l'ajout de traits de prédiction (ADDPRED + SELEC) correspond à une claire tendance vers une amélioration significative en termes d'exactitude par rapport à un entraînement sur les seules données naturelles (p-valeur de $0{,}033$ avec un effet de taille important de $0{,}756$ et une corrélation forte, $0{,}842$). Les scores de F1 pour toutes les classes sont améliorés : $20{,}8$ % pour *contraste*, $51{,}0$ % pour *résultat*, $31{,}2$ % pour *continuation* et $53{,}9$ % pour *explication*. Deux autres systèmes obtiennent un score d'exactitude supérieur à 40 % : NATW + SELEC ($41{,}3$ %, avec une tendance vers une amélioration significative[9]) et UNION + SELEC ($40{,}1$ %, non significativement meilleur). Notons que le système ADDPRED correspond à la meilleur *baseline* dans (Daumé III et Marcu, 2006), ce qui tend à montrer que prendre en compte la différence de distribution entre nos données sous l'angle de l'adaptation de domaine est pertinent.

De manière générale, la phase de sélection permet d'autoriser un poids plus fort sur les informations provenant des données artificielles. Pour le système LININT + SELEC, les meilleurs résultats sont obtenus avec une influence quasiment égale des deux modèles. De même, pour NATW + SELEC, la moyenne des coefficients choisis est beaucoup plus basse, et elle augmente largement pour ARTSUB + SELEC, autorisant des sous-ensembles plus larges, avec cependant un fort éparpillement des valeurs. Ces considérations accréditent malgré tout l'hypothèse selon laquelle la sélection améliore la qualité du corpus artificiel. Au niveau des seuils choisis, la moyenne sur toutes les expériences est aux alentours de $0{,}7$, avec un écart-type variable selon les systèmes mais supérieur à $0{,}1$. C'est un écart assez important, cet hyperparamètre semble relativement mal estimé par notre validation croisée interne et nécessiterait de futures expériences, où l'on pourra réduire l'espace de recherche.

La sélection des exemples améliore l'identification des relations et conduit à un système améliorant significativement l'exactitude de NATONLY montrant que les données artificielles lorsque intégrées de façon adéquate peuvent améliorer l'identification des relations implicites, notamment lorsque leur influence est faible, le modèle étant guidé vers la bonne distribution.

9. p-valeur de $0,077$, effet de taille large de $0{,}68$ et importante corrélation $0{,}67$.

À la constitution des corpus avec sélection, nous avons observé qu'avec la croissance du seuil, nous conservions toujours plus d'exemples pour *résultat*. Dès le seuil $0{,}4$, nous disposons d'environ $3\ 900$ exemples de plus pour cette relation, alors que *contraste* devient sous-représenté. Cette observation montre que le bruit n'est probablement pas la seule façon d'expliquer les résultats puisque la relation améliorée par les données artificielles est celle pour laquelle le modèle artificiel est le moins confiant alors que la relation dont les résultats sont les plus dégradés est celle pour laquelle le modèle est le plus confiant.

7. Conclusion et perspectives

Nous avons développé la première série de systèmes d'identification des relations discursives implicites pour le français. Ces relations sont difficiles à identifier en raison du manque d'indices forts. Dans les études sur l'anglais, les performances sont basses malgré les indices complexes utilisés, probablement par manque de données. Pour pallier ce problème, plus crucial encore en français, nous avons utilisé des données annotées automatiquement en relation à partir d'exemples explicites. S'il semble bien que la stratégie ait une certaine pertinence, au sens où un connecteur peut être redondant avec son contexte, ces nouvelles données ne permettent pas une généralisation qui convient aux données implicites car elles sont de distributions différentes. Nous avons donc testé des méthodes inspirées de l'adaptation de domaine pour combiner ces données en ajoutant une étape de sélection automatique des exemples artificiels pour gérer le bruit induit par leur création. Cette stratégie nous permet d'aboutir à des améliorations significatives par rapport au modèle n'utilisant que les données naturelles. Les meilleurs systèmes utilisent la sélection d'exemples, le meilleur repose sur l'ajout des traits de prédiction ($41{,}7$ % d'exactitude), le deuxième meilleur correspond à la pondération des données manuelles ($41{,}3$ %).

Dans un premier temps, il faudrait porter ces méthodes sur les données anglaises afin de valider par comparaison cette première approche et de vérifier s'il est possible d'obtenir des résultats du même ordre sur une autre langue. Ensuite, comme les méthodes de combinaison et de sélection simples utilisées ici parviennent à des résultats encourageants, on peut espérer que des méthodes plus sophistiquées pourraient conduire à des améliorations plus importantes. Nous envisageons de tester d'autres méthodes utilisées en adaptation de domaine, comme l'approche par régularisation proposée par Chelba et Acero (2006), ou celle fondée sur une modification de l'espace de représentation de Daumé III (2007). Enfin, une étude des données explicites permettrait d'augmenter la taille du corpus artificiel. Cela pourrait aussi amener à améliorer sa qualité en sélectionnant des connecteurs pour lesquels cette méthode est plus ou moins efficace. Il faudrait aussi identifier les relations pour lesquelles cette méthode n'est pas possible, soit parce que ces relations n'apparaissent jamais sans connecteur (comme les contrastes de type violation d'attente) soit, au contraire, parce qu'elles n'apparaissent jamais avec un connecteur (comme la relation d'encadrement

dans ANNODIS). Cette étude permettrait aussi probablement de trouver les traits les plus informatifs dans une optique de combinaison des données.

8. Bibliographie

Afantenos S., Asher N., Benamara F., Bras M., Fabre C., Ho-Dac L.-M., Le Draoulec A., Muller P., Pery-Woodley M.-P., Prévot L., Rebeyrolles J., Tanguy L., Vergez-Couret M., Vieu L., « An empirical resource for discovering cognitive principles of discourse organisation : the ANNODIS corpus », *in* N. Calzolari (ed.), *Proceedings of LREC*, 2012.

Asher N., Lascarides A., *Logics of Conversation*, Cambridge University Press, 2003.

Asr F., Demberg V., « Implicitness of discourse relations », *Proceedings of COLING*, 2012.

Berger A. L., Pietra V. J. D., Pietra S. A. D., « A maximum entropy approach to natural language processing », *Computational linguistics*, 1996.

Blair-Goldensohn S., McKeown K. R., Rambow O. C., « Building and refining rhetorical-semantic relation models », *Proceedings of NAACL HLT*, 2007.

Candito M., Nivre J., Denis P., Anguiano E. H., « Benchmarking of statistical dependency parsers for French », *Proceedings of ICCL (posters)*, 2010.

Carlson L., Marcu D., Okurowski M. E., « Building a discourse-tagged corpus in the framework of rhetorical structure theory », *Proceedings of SIGdial*, 2001.

Chelba C., Acero A., « Adaptation of maximum entropy capitalizer : Little data can help a lot », *Computer Speech & Language*, 2006.

Daumé III H., « Frustratingly Easy Domain Adaptation », *Proceedings of ACL*, 2007.

Daumé III H., Kumar A., Saha A., « Frustratingly Easy Semi-supervised Domain Adaptation », *Proceedings of DANLP*, 2010.

Daumé III H., Marcu D., « Domain adaptation for statistical classifiers », *Journal of Artificial Intelligence Research*, 2006.

de Winter J., « Using the Student's t-test with extremely small sample sizes », *Practical Assessment, Research & Evaluation*, 2013.

Denis P., Sagot B., « Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort », *Proceedings of PACLIC*, 2009.

Hernault H., Prendinger H., duVerle D. A., Ishizuka M., « HILDA : A Discourse Parser Using Support Vector Machine Classification », *Dialogue and Discourse*, 2010.

Jiang J., « A Literature Survey on Domain Adaptation of Statistical Classifiers », , Available from : http ://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf, 2008.

Lin Z., Kan M.-Y., Ng H. T., « Recognizing Implicit Discourse Relations in the Penn Discourse Treebank », *Proceedings of EMNLP*, vol. 1, 2009.

Lin Z., Ng H. T., Kan M. Y., « A PDTB-styled end-to-end discourse parser », *CoRR*, 2010.

Mann W. C., Thompson S. A., « Rhetorical Structure Theory : Toward a functional theory of text organization », *Text*, 1988.

Marcu D., Echihabi A., « An Unsupervised Approach to Recognizing Discourse Relations », *Proceedings of ACL*, 2002.

Moeschler J., « Connecteurs, encodage conceptuel et encodage procédural », *Cahiers de Linguistique Française*, 2002.

Moreno-Torres J. G., Raeder T., Alaiz-Rodríguez R., Chawla N. V., Herrera F., « A unifying view on dataset shift in classification », *Pattern Recognition*, 2012.

Muller P., Afantenos S., Denis P., Asher N., « Constrained decoding for text-level discourse parsing », *Proceedings of COLING*, 2012.

Park J., Cardie C., « Improving Implicit Discourse Relation Recognition Through Feature Set Optimization », *Proceedings of SIGdial*, 2012.

Pitler E., Louis A., Nenkova A., « Automatic sense prediction for implicit discourse relations in text », *Proceedings of ACL-IJCNLP*, 2009.

Pitler E., Nenkova A., « Using Syntax to Disambiguate Explicit Discourse Connectives in Text », *Proceedings of ACL-IJCNLP*, 2009.

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B., « The penn discourse treebank 2.0 », *Proceedings of LREC*, 2008.

Roze C., Vers une algèbre des relations de discours, These, Université Paris 7, May, 2013.

Roze C., Danlos L., Muller P., « LEXCONN : A French Lexicon of Discourse Connectives », *Discours*, 2012.

Sagae K., « Analysis of Discourse Structure with Syntactic Dependencies and Data-Driven Shift-Reduce Parsing », *Proceedings of IWPT*, 2009.

Sanders T., « Coherence, causality and cognitive complexity in discourse », *Proceedings of the Symposium on the Exploration and Modelling of Meaning*, 2005.

Scheffer T., Error Estimation and Model Selection, PhD thesis, Technischen Universitet Berlin, School of Computer Science, 1999.

Sogaard A., *Semi-supervised learning and domain adaptation in natural language processing*, Morgan & Claypool, 2013.

Soria C., Ferrari G., « Lexical marking of discourse relations-some experimental findings », *Proceedings of ACL (Workshop on Discourse Relations and Discourse Markers)*, 1998.

Sporleder C., Lascarides A., « Exploiting linguistic cues to classify rhetorical relations », *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 2007.

Sporleder C., Lascarides A., « Using Automatically Labelled Examples to Classify Rhetorical Relations, An Assessment », *Natural Language Engineering*, 2008.

Varma S., Simon R., « Bias in error estimation when using cross-validation for model selection », *BMC bioinformatics*, vol. 7, n° 1, p. 91, 2006.

Wang X., Li S., Li J., Li W., « Implicit Discourse Relation Recognition by Selecting Typical Training Examples », *Proceedings of COLING*, 2012.

Webber B., « D-LTAG : extending lexicalized TAG to discourse », *Cognitive Science*, 2004.