Identification automatique des relations discursives « implicites » à partir de données annotées et de corpus bruts

Chloé Braud[1] Pascal Denis[2]
(1) ALPAGE, INRIA Paris-Rocquencourt & Université Paris Diderot
(2) MAGNET, INRIA Lille Nord-Europe
chloe.braud@inria.fr, pascal.denis@inria.fr

RÉSUMÉ

Cet article présente un système d'identification des relations discursives dites « implicites » (à savoir, non explicitement marquées par un connecteur) pour le français. Etant donné le faible volume de données annotées disponibles, notre système s'appuie sur des données étiquetées automatiquement en supprimant les connecteurs non ambigus pris comme annotation d'une relation, une méthode introduite par (**?**). Comme l'ont montré (**?**) pour l'anglais, cette approche ne généralise pas très bien aux exemples de relations implicites tels qu'annotés par des humains. Nous arrivons au même constat pour le français et, partant du principe que le problème vient d'une différence de distribution entre les deux types de données, nous proposons une série de méthodes assez simples, inspirées par l'adaptation de domaine, qui visent à combiner efficacement données annotées et données artificielles. Nous évaluons empiriquement les différentes approches sur le corpus ANNODIS : nos meilleurs résultats sont de l'ordre de 45.6% d'exactitude, avec un gain significatif de 5.9% par rapport à un système n'utilisant que les données annotées manuellement.

ABSTRACT

Automatically identifying implicit discourse relations using annotated data and raw corpora

This paper presents a system for identifying « implicit » discourse relations (that is, relations that are not marked by a discourse connective). Given the little amount of available annotated data for this task, our system also resorts to additional automatically labeled data wherein unambiguous connectives have been suppressed and used as relation labels, a method introduced by (**?**). As shown by (**?**) for English, this approach doesn't generalize well to implicit relations as annotated by humans. We show that the same conclusion applies to French due to important distribution differences between the two types of data. In consequence, we propose various simple methods, all inspired from work on domain adaptation, with the aim of better combining annotated data and artificial data. We evaluate these methods through various experiments carried out on the ANNODIS corpus : our best system reaches a labeling accuracy of 45.6%, corresponding to a 5.9% significant gain over a system solely trained on manually labeled data.

MOTS-CLÉS : analyse du discours, relations discursives, apprentissage automatique.

KEYWORDS: discourse analysis, discourse relations, machine learning.

1 Introduction

L'analyse discursive rend compte de la cohérence d'un texte en liant, par des relations discursives, les propositions qui le constituent. En dépit de différences, les principales théories du discours, telles que la Rhetorical Structure Theory (RST) (**?**) et la Segmented Discourse Representation Theory (SDRT) (**?**), s'accordent sur les étapes d'analyse : segmentation en unités élémentaires de discours (EDU), attachement des EDU, identification des relations entre EDU, puis récursivement les paires attachées sont liées à des segments simples ou complexes pour aboutir à une structure couvrant le document. Ainsi on peut associer au discours 1.1 [1] segmenté en trois EDU la structure entre accolades. Les deux premiers segments sont liés par un *contrast* et le segment complexe ainsi constitué est argument d'une relation de *continuation*. Un système dérivant automatiquement cette structure permettrait d'améliorer d'autres systèmes de TAL ou de RI car la structure du discours contraint les référents des anaphores, révèle la structure thématique d'un texte et l'ordonnancement temporel des événements : dans 1.2, les phrases *a* et *b* sont liées par une relation de type *explanation*, *b* explique *a*, qui implique (loi de cause à effet) l'ordre des événements, *b* avant *a*.

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

Exemple 1.2 {[*Juliette est tombée.*]$_a$ [*Marion l'a poussée.*]$_b$}$_{explanation}$

Grâce aux corpus annotés comme le PDTB [2] ou le RST DT [3] des systèmes automatiques ont été développés pour l'anglais sur la tâche complète ou seulement les sous-tâches (notamment la phase d'identification des relations). A partir du corpus RST DT, (**?**) et (**?**) ont développé des systèmes complets avec des scores de f-mesure respectifs de 44.5 et 47.3 donc des performances encore modestes. Sur le PDTB, (**?**) construit un système complet obtenant 46.8 de f-mesure.

Le PDTB permet de séparer l'étude des exemples avec ou sans connecteur discursif déclenchant la relation. Lorsqu'un tel marqueur est présent, la relation est dite explicite (ou marquée ou lexicalisée) : ainsi, *mais* lexicalise la relation de *contrast* dans 1.1. Sinon, elle est implicite, comme la relation causale dans 1.2. Les différentes études menées sur le PDTB montrent que l'identification des relations implicites est considérablement plus difficile que celle des relations explicites. Ainsi, (**?**) obtiennent une f-mesure qui dépasse les 80 pour les explicites, mais de seulement 39.63 pour les implicites. Sur un jeu de relations plus petit, (**?**) rapportent une exactitude de 94% sur les explicites alors que (**?**) de 60 sur les implicites. Sur des données tirées du RST DT, avec 5 relations, (**?**) obtiennent des scores de l'ordre de 40% d'exactitude. Pour le français, il n'existe pas de corpus annoté en connecteur, donc aucune étude séparant le cas des implicites du cas général : (**?**) ont développé un système complet qui obtient une exactitude de 44.8 pour 17 relations et de 65.5 pour ces relations regroupées en 4 classes (ANNODIS, 3143 exemples). On peut supposer que, comme pour l'anglais, ce sont les relations implicites qui dégradent les performances du système.

Malheureusement, les corpus discursifs disponibles sont encore très petits (surtout pour le français). En vue de pallier le manque d'annotations humaines, (**?**) proposent d'utiliser des exemples

1. Tiré du corpus français ANNODIS, (**?**) document WK_-_hulotte.
2. Penn Discourse Treebank, (**?**)
3. RST Discourse Treebank, (**?**)

annotés automatiquement grâce aux connecteurs comme données implicites supplémentaires. Cette étude et celles qui l'ont suivie, notamment (**?**), utilisaient ces nouvelles données artificielles comme *seules* données d'entraînement et obtenaient de basses performances. Le problème repose sur une différence de distribution entre les deux types de données, qu'il est possible de prendre en compte afin d'améliorer l'identification des relations implicites. A cette fin, nous proposons et évaluons différentes méthodes visant à créer un nouveau modèle enrichi par les nouvelles données mais guidé vers la distribution des données manuelles. Nous nous inspirons des méthodes utilisées en adaptation de domaine décrites dans (**?**). Notre contribution se situe au niveau du développement d'un système d'identification des relations discursives implicites pour le français et de l'étude de stratégies d'utilisation de données de distributions différentes en TAL.

Nous présentons dans la partie suivante un rapide état de l'art sur les expériences déjà menées sur l'identification des relations de discours avec données artificielles, afin d'en montrer les limites et de proposer une nouvelle stratégie. La section 3 est consacrée aux données et la section 4 au modèle utilisé. La section 5 regroupe les expériences menées et l'analyse des résultats. Enfin, nous finirons par les perspectives ouvertes par ces expériences dans la section 6.

2 Utilisation des données générées automatiquement

Les obstacles associés à l'identification des relations implicites résident, d'une part, dans l'absence d'indicateur fiable (comme le connecteur pour les relations explicites) et, d'autre part, dans le manque de données pour entraîner des classifieurs performants. Néanmoins, on dispose de données quasiment annotées en grande quantité : celles contenant un connecteur discursif non ambigu, c'est-à-dire ne déclenchant qu'une seule relation (p.ex., *parce que* déclenche nécessairement une relation de type *explanation*). Ce constat a amené (**?**) à proposer d'utiliser ces exemples pour l'identification des implicites. Plus précisément, on génère de nouvelles données annotées à partir d'un corpus brut : des exemples sont extraits sur la présence d'une forme de connecteur discursif non ambigu, filtrés pour éliminer les cas d'emploi non discursif de la forme, puis le connecteur est supprimé pour empêcher le modèle de se baser sur cet indice non ambigu. On crée ainsi des données implicites annotées en relation de discours mais des données qui n'ont jamais été produites, non naturelles d'où le terme de données artificielles. A titre d'illustration, considérons la paire de phrases suivante tirée du corpus Est Républicain (2.1) : dans ce cas, le connecteur *cela dit* est supprimé et on génère un exemple de relation de *contrast* entre les deux syntagmes arguments a et b.

Exemple 2.1 $[$*Elle était très comique, très drôle.*$]_a$ Cela_dit $[$*,le drame n' était jamais loin.*$]_b$

L'idée est finalement de s'appuyer sur ces données artificielles pour construire un modèle d'identification des relations pour des données naturelles implicites, on a donc des données de type différent : implicites *versus* explicites et naturelles *versus* artificielles.

Dans les études précédentes basées sur ce principe les données artificielles sont utilisées comme seules données d'entraînement ce qui conduit à des performances basses, juste au-dessus de la chance pour (**?**) avec 25.8% d'exactitude contre 40.3 en utilisant les seules données manuelles (1051 exemples manuels, 72000 artificiels, 5 relations). (**?**) cherchent à tester l'impact de la qualité du corpus artificiel en améliorant l'extraction des données grâce à une segmentation en

topics ou des informations syntaxiques. Ils semblent améliorer légèrement les performances mais une comparaison est difficile puisqu'ils ne testent que des classifieurs binaires et 2 relations. L'idée de base de (**?**) résidait dans l'extraction de paires de mots de type antonymes ou hypéronymes pouvant révéler une relation mais dont le lien n'est pas forcément recensé dans des ressources comme WordNet. (**?**) montrent que l'utilisation des paires de mots extraites d'un corpus artificiel comme trait supplémentaire n'améliore pas les performances d'un système d'identification des relations implicites. Mais l'étude de (**?**) utilisant d'autres types de traits indique que le problème ne réside pas ou pas uniquement dans le choix des traits.

Ces résultats montrent qu'un modèle entraîné sur les données artificielles ne généralise pas bien aux données manuelles. Pourtant en regardant des exemples de type artificiel, il semble que dans certains cas on aurait pu produire les arguments sans le connecteur. De plus, les résultats de (**?**) demeurent supérieurs à la chance (en considèrant la chance à 20%), donc ces données ne sont pas complètement différentes des données de test. Nous cherchons ici à prendre en compte cette différence de distribution qui rapproche le problème de ceux traités en adaptation de domaine.

2.1 Problème : différence de distribution entre les données

Pour que cette stratégie fonctionne, il faut nécessairement faire l'hypothèse d'une certaine redondance du connecteur par rapport à son contexte : il doit rester suffisamment d'information après sa suppression pour que la relation reste identifiable. Une étude psycho-linguistique menée sur l'italien (**?**) et les conclusions de (**?**) semblent montrer que c'est le cas dans une partie des données. Cette étude reste à faire pour le français, et l'approfondir pourrait permettre d'améliorer la qualité du corpus artificiel en déterminant par exemple si cette redondance est différente selon les relations et les connecteurs.

Plus généralement, en apprentissage on fait l'hypothèse que données d'entraînement et de test sont identiquement et indépendamment distribuées (*données i.i.d.*). Or il nous semble que justement la stratégie proposée par (**?**) pose le problème d'un apprentissage avec des données non identiquement distribuées. On a deux ensembles de données qui se ressemblent (même ensemble d'étiquettes, les exemples sont des segments de texte) mais qui sont néanmoins distribués différemment, et ce, pour deux raisons au moins. 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 "vrais" exemples implicites. La différence porte tant sur la distribution des labels (des relations) que sur l'assocation entre labels (relations) et inputs (paires des segments) à classer. En outre, la suppression du connecteur ajoute probablement une forme de bruit en cas d'erreur d'étiquetage contrairement aux données manuelles correctement étiquetées.

D'autre part, les données artificielles se distinguent aussi des données manuelles en termes des segments. Ainsi, la segmentation des premières est basée sur des heuristiques (p.ex., les arguments ne peuvent être que deux phrases adjacentes ou deux propositions couvrant une phrase). Dans les données manuelles, en revanche, on a des arguments contigus ou non, propositionnels, phrastiques ou multi-phrastiques dont les frontières ont été déterminées par des annotateurs humains. Ceci induit une différence de distribution au niveau des objets à classer et une forme de bruit en cas d'erreur de segmentation due à ces hypothèses simplificatrices ou à une erreur d'heuristique.

On peut se rendre compte de cette différence de distribution sur l'association entre labels et inputs en considérant certaines caractéristiques des données. La répartition entre exemples inter-phrastiques et intra-phrastiques (la relation s'établit entre deux phrases ou deux segments à l'intérieur d'une phrase) est ainsi similaire pour *contrast* (57.1% d'inter-phrastiques dans les deux types de données), proche pour *result* (45.7% d'inter-phrastiques dans les données manuelles, 39.8% dans les artificielles) mais très différente pour *continuation* (70.0% d'inter-phrastiques dans les manuelles, 96.5% dans les artificielles), et pour *explanation* (21.4% dans les manuelles, 53.0% dans les artificielles).

Ne pas prendre en compte ces différences de distribution conduit à de basses performances, nous avons donc cherché à les gérer en testant différentes stratégies avec un point commun : chercher à guider le modèle vers la distribution des données manuelles.

2.2 Modèles testés

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

De nombreux travaux s'attachant au problème de données non identiquement distribuées concernent l'adaptation de domaine. Nous nous sommes donc inspirés des méthodes utilisées dans ce cadre, même si notre problème diffère au sens où nous n'avons qu'un seul domaine et des données bruitées. Ainsi, nous avons testé une série de systèmes utilisés pour l'adapation de domaine par (**?**), qui sont très simples à mettre en oeuvre et obtiennent néanmoins de bonnes performances sur différentes tâches, ainsi que quelques solutions dérivées. Dans un second temps, nous avons ajouté une étape de sélection d'exemples, afin de choisir parmi les exemples artificiels ceux qui seraient susceptibles d'améliorer les performances.

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. La première stratégie de combinaison de données 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. Une stratégie dérivée (ARTIKMAN) consiste à prendre, non pas l'intégralité des données artificielles, mais des sous-ensembles aléatoires de ces données, en addition des données manuelles. Cette méthode est un peu plus subtile dans la mesure où on peut faire varier la proportion des exemples artificiels par rapport aux exemples manuels. Enfin, la troisième méthode du premier type (MANKMAN) garde cette fois la totalité des données artificielles mais pondère (ou duplique) les exemples manuels de manière à éviter un déséquilibre trop grand au profit des données artificielles.

Dans le second type de méthodes, on trouve tout d'abord une méthode (ADDPRED) qui consiste à utiliser les prédictions d'un modèle entraîné sur les données artificielles (à savoir les données "source") comme descripteur dans le modèle entraîné sur les données manuelles (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 (**?**). Une variation de cette méthode (ADDPROBA)

utilise en plus le score de confiance (p.ex., la probabilité) du modèle artificiel comme descripteur supplémentaire dans le modèle manuel. Une troisième méthode (INIT) vise à initialiser les paramètres du modèle entraîné sur les données manuelles avec ceux du modèle utilisant les données artificielles. Enfin, la dernière méthode (INTERPLIN) se base sur une interpolation linéaire de deux modèles préalablement entraînés sur chacun des ensembles de données.

Nous avons aussi testé toutes ces stratégies en ajoutant une étape de sélection automatique d'exemples artificiels. La méthode utilisée est naïve puisqu'elle se base simplement sur la probabilité du label prédit : on teste différents seuils sur ces probabilités en ajoutant à chaque fois les seuls exemples prédits avec une probabilité supérieure au seuil. Cette sélection vise à écarter des données bruitées, en explorant finalement l'une des voies proposées par (**?**) et développée d'une autre manière par (**?**), à savoir améliorer la qualité du corpus artificiel.

Les performances de tous ces systèmes seront comparées à celles des systèmes entraînés séparément sur les deux ensembles de données dans la section 5.

3 Données

Nous avons choisi de nous restreindre à 4 relations : *contrast*, *result*, *continuation* et *explanation*. Ces relations sont annotées dans le corpus français utilisé et correspondent à des exemples implicites et explicites. De plus ce sont 4 des 5 relations (*summary* n'est pas annotée dans ANNODIS) utilisées dans (**?**), ce qui nous permet une comparaison mais non directe puisque la langue et le corpus sont différents. Dans nos données manuelles, nous avons fusionné les méta-relations avec les relations correspondantes avec l'hypothèse qu'elles mettaient en jeu le même genre d'indices et de constructions.

3.1 Le corpus ANNODIS

Le projet ANNODIS (**?**) vise la construction d'un corpus annoté en discours pour le français suivant le cadre SDRT. La version du corpus utilisée (en date du 15/11/2012) comporte 86 documents provenant de l'Est Républicain et de Wikipedia. 3339 exemples sont annotés avec 17 relations rhétoriques. 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, contiguës ou non. Les connecteurs discursifs ne sont pas annotés.

Le corpus a subi une série de pré-traitements. Le MElt tagger (**?**) fournit un étiquetage en catégorie morpho-syntaxique, une lemmatisation, des indications morphologiques (temps, personne, genre, nombre). Le MSTParser (**?**) 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 (**?**) et étendu en 2012 aux connecteurs introduisant des syntagmes nominaux. Nous utilisons une méthode simple en projetant simplement les connecteurs sur les données (sauf *à* jugé trop ambigu), sans restriction de position, le but étant d'être certain de ne conserver que des implicites au risque d'en perdre certains. Sur les 1108 exemples disponibles pour les 4 relations nous disposons de 494 exemples implicites ; la distribution des exemples par relation est résumée dans le tableau 1.

Relation	Exemples explicites	Exemples implicites	Total
contrast	100	42	142
result	52	110	162
continuation	404	272	676
explanation	58	70	128
all	614	494	1108

TABLE 1 – Corpus ANNODIS : nombre d'exemples explicites et implicites par relation

3.2 Le corpus généré automatiquement

Nous avons utilisé 100 connecteurs du *LexConn* de (**?**) pour identifier des formes de connecteur ne pouvant déclencher qu'une relation parmi les 4 choisies dans le corpus composé d'articles de l'Est Républicain (9M de phrases), avec les mêmes traitements que pour ANNODIS. Les exemples sont ensuite filtrés pour éliminer les emplois non discursifs en tenant compte de la position du connecteur et de la ponctuation et en s'aidant des indications de *LexConn*. L'identification des arguments d'un connecteur est 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, au plus 2 EDU par phrase.

Cette méthode simple permet de générer rapidement de gros volumes de données : au total, nous avons pu extraire 392260 exemples (voir tableau 2). Lorsque plusieurs connecteurs étaient détectés, nous avons généré au maximum deux exemples si l'un est intra-phrastique et l'autre inter-phrastique pour éviter de générer un exemple pour un connecteur et son modifieur. Nous avons équilibré le corpus en relation 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 manuelles et artificielles : *continuation* la plus représentée dans les manuelles devient la moins représentée dans les artificielles. Ceci est dû à la forte ambiguïté des connecteurs de cette relation qui nous ont forcé 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.

Relation	Disponible	Entraînement	Développement	Test	Total
contrast	252 793	23 409	2 926	2 926	29 261
result	50 297	23 409	2 926	2 926	29 261
continuation	29 261	23 409	2 926	2 926	29 261
explanation	59 909	23 409	2 926	2 926	29 261
all	392 260	93 636	11 704	11 704	117 044

TABLE 2 – Corpus artificiel : nombre d'exemples par relation

4 Modèle et jeu de traits

Pour cette étude, nous avons employé un modèle de classification discriminant par régression logistique (ou maximum d'entropie). Ce choix est basé sur le fait que ce type de modèles donne de

bonnes performances pour différents problèmes de TAL et a été implanté dans différentes librairies librement disponibles. Le principe de cet algorithme est d'apprendre un jeu de paramètres qui maximise la log-vraisemblance des données fournies à l'apprentissage (voir (**?**)). Un attrait important de ces modèles, par rapport à des modèles génératifs, est de permettre l'ajout de nombreux descripteurs potentiellement redondants sans faire d'hypothèses d'indépendance.

Notre jeu de traits se base 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, (**?**) ne se base que sur la co-occurrence de lemmes dans les segments. (**?**) montrent que la prise en compte de différents types de traits linguistiquement motivés améliore les performances. (**?**) utilisent des traits variés dont des bigrammes de lemmes mais sans traits syntaxiques. Nous avons testé des traits lexico-syntaxiques 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 car les ressources nécessaires n'existent pas pour le français.

Certains traits sont calculés pour chaque argument :

1. Indice de complexité syntaxique : nombre de syntagmes nominaux, verbaux, prépositionnels, adjectivaux, adverbiaux (valeur continue)

2. Information sur la tête d'un argument :
   - Lemme d'éléments négatifs sur la tête (booléen)
   - Information temporelle/aspectuelle : nombre de fois où un lemme de fonction auxiliaire dépendant de la tête apparaît (valeur continue), temps, personne, nombre de l'auxiliaire (booléen)
   - Informations sur les dépendants de la tête : présence d'un objet, par-objet, modifieur ou dépendant prépositionnel de la tête, du sujet ou de l'objet (booléen) ; catégorie morpho-syntaxique 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 morpho-syntaxique précise (par exemple "VPP") et simplifiée (respectivement "V") (booléen)

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

1. Trait de position : si l'exemple est inter ou intra-phrastique (booléen)

2. Indice de continuité thématique : chevauchement en lemmes et en lemmes de catégorie ouverte (continue)

3. Information sur les têtes des arguments :
   - Paire des temps des têtes verbales (booléen)
   - Paire des nombres des têtes (booléen)

On notera 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 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 2, peut nous permettre de dépasser les performances obtenues en s'entraînant sur des données manuelles présentes seulement en faible quantité.

Les expériences sont réalisées avec l'implémentation de l'algorithme par maximum d'entropie fourni dans la librairie MegaM [4] en version multi-classe avec au maximum 100 itérations. On effectue une validation croisée en 10 sous-ensembles sur un corpus des données manuelles équilibré à 70 exemples 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. Comme dans les études précédentes, les performances sont données en termes d'exactitude globale sur l'ensemble des relations, des scores ventilés de F1 par relation sont également fournis. La significativité statistique des écarts de performance est évaluée avec un Wilcoxon signed-rank test (avec une p-valeur < 0.05).

5.1 Modèles de base

Dans un premier temps, nous construisons deux modèles distincts, l'un à partir des seules données manuelles (MANONLY, 252 exemples), l'autre des seules données artificielles (ARTIONLY, 93636 exemples d'entraînement). Notre modèle MANONLY obtient une exactitude de 39.7, avec des scores de f-mesure par relation compris entre 13.3 pour *contrast* et 49.0 pour *result* (voir table 3). La relation *contrast* 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 ARTIONLY obtient une exactitude de 47.8 lorsqu'évalué sur le même type de données (11704 exemples de test), mais de 23.0 lorsqu'évalué sur les données manuelles (voir table 3). 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 2. De manière générale, on observe des dégradations par rapport à MANONLY pour l'identification de *result*, *explanation* et *continuation* (voir table 3). Par contre l'identification de *contrast* présente une amélioration, obtenant 23.2 de f-mesure avec 11 exemples correctement identifiés contre 5 précédemment.

	MANONLY	ARTIONLY	
Données de test	Manuelles	Manuelles	Artificielles
Exactitude	39.7	23.0	47.8
contrast	13.3	23.2	38.3
result	49.0	15.7	57.4
continuation	39.7	32.1	54.3
explanation	43.8	22.4	37.5

TABLE 3 – Modèles de base, exactitude du système et f-mesure par relation

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

5.2 Modèles avec combinaisons de données

Dans cette section, nous présentons les résultats des systèmes qui exploitent à la fois les données manuelles 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 hyper-paramètres. Ainsi, pour la pondération des exemples manuels nous testons différents coefficients de pondération c avec $c \in [0.5; 2000]$ avec un incrément de 10 jusqu'à 100, de 50 jusqu'à 1000 et de 500 jusqu'à 2000. Pour l'ajout de sous-ensembles des données artificielles, on ajoute à chaque fois k exemples parmi ces données avec $k \in [0.1; 600]$ avec un incrément de 10 jusqu'à 100 et de 50 jusqu'à 600. Enfin, pour l'interpolation linéaire des modèles, on construit un nouveau modèle en pondérant le modèle artificiel avec $\alpha \in [0.1; 0.9]$ avec des incréments de 0.1.

De manière générale, l'ensemble de ces systèmes avec les bons hyper-paramètres conduit à des résultats au moins équivalents et parfois supérieurs en exactitude par rapport à MANONLY. Si la tendance générale est donc plutôt d'une hausse des performances, aucune des différences observées à ce stade ne semble cependant être statistiquement significative. Les scores des systèmes présentant les résultats les plus pertinents sont repris dans la table 4.

	MANONLY	ARTIONLY	UNION	MANKMAN		ARTIKMAN	ADDPRED	ADDPROBA	INIT	INTERPLIN		
Paramètre	-	-	-	100	400	0.2	-	-	-	0.2	0.5	0.8
Exactitude	39.7	23.0	24.2	34.9	41.7	39.7	**42.9**	39.3	39.3	39.7	38.5	35.3
contrast	13.3	23.2	21.1	**32.4**	19.7	16.7	22.9	24.0	11.9	13.2	16.2	29.6
result	49.0	15.7	16.4	28.0	39.2	44.9	47.4	45.8	46.3	47.0	39.5	24.8
continuation	39.7	32.1	38.5	45.8	**48.7**	39.4	37.3	35.4	38.9	40.6	39.2	44.2
explanation	43.8	22.4	21.7	31.7	47.7	46.1	**52.8**	43.8	45.4	45.4	48.6	40.3

TABLE 4 – Modèles sans sélection d'exemples, exactitude du système et f-mesure par relation

La seule configuration qui mène à des résultats négatifs est l'union simple des corpus d'entraînement (UNION). Ce système obtient 24.2 d'exactitude donc de l'ordre d'un entraînement sur les seules données artificielles. Ces résultats ne sont pas surprenants, les données manuelles environ 372 fois moins nombreuses que les artificielles se retrouvent noyées dans les données artificielles.

Les expériences de combinaison des données, ajout de sous-ensembles aléatoires des données artificielles (ARTIKMAN) et pondération des exemples manuels (MANKMAN), ont des tendances inverses. Avec ARTIKMAN, l'exactitude diminue lorsque le coefficient augmente (de 40.1 à 21.8) et atteint ou dépasse le modèle manuel avec les coefficients 0.1 et 0.2, donc une influence très faible du modèle artificiel. Avec MANKMAN, l'exactitude augmente avec la croissance du coefficient (de 24.2 à 40.9) et dépasse 39.7 à partir du coefficient 400 équivalent à un corpus manuel d'environ 24000 exemples par relation soit du même ordre que le nombre d'exemples artificiels.

Les expériences où la prise en compte des données artificielles passe par l'ajout de traits donnent les meilleurs résultats avec une exactitude de 42.9 pour le modèle qui intègre les prédictions du modèle artificiel comme descripteur (ADDPRED). Le second modèle, qui exploite en plus les probabilités (ADDPROBA) mène quant à lui à une légère diminution ce qui suggère que les traits de probabilité dégradent les performances.

Quant aux expériences de combinaison des modèles, l'initialisation du modèle manuel par

l'artificiel (INIT) conduit à un système d'exactitude 39.3, et l'interpolation linéaire (INTERPLIN) correspond à une décroissance de l'exactitude suivant l'augmentation du coefficient α sur le modèle artificiel (voir table 4), avec cependant un saut important entre $\alpha = 0.8$ et $\alpha = 0.9$ (exactitude de 28.2) en lien avec une forte dégradation de l'identification de *explanation*.

Au niveau des scores par relation, ces systèmes ont des effets différents. Une influence forte du modèle artificiel permet une amélioration importante pour *contrast* et une dégradation forte pour *result* et *explanation* par rapport à MANONLY. Ces phénomènes sont visibles avec ARTIONLY mais aussi avec INTERPLIN : la f-mesure de *contrast* augmente avec α tandis que celle de *result* et de *explanation* diminue (voir table 4). Avec une influence similaire des deux types de données (INTERPLIN $\alpha = 0.5$ ou MANKMAN coefficient 400), la chute pour *result* est moins importante et on améliore l'identification de *explanation* (voir table 4). Pour *continuation*, il faut une influence des données manuelles inférieure à celle des données artificielles pour observer une amélioration (voir table 4, INTERPLIN $\alpha = 0.8$). Le système ADDPRED permet notamment une amélioration forte de la f-mesure de *explanation*. On n'obtient pas d'amélioration pour *result*.

Les méthodes de combinaison aboutissent à des systèmes d'exactitude similaire voire supérieure à MANONLY et à des améliorations pour l'identification des relations sauf *result*. La relation *contrast* profite peut-être de données artificielles moins bruitées : la majorité des exemples (plus de 75%) sont extraits à partir de *mais*, forme toujours en emploi discursif dont les arguments sont dans l'ordre canonique, argument1+connecteur+argument2. Pour *explanation*, la majorité des données (77.5%) est extraite à partir de formes déclenchant la méta-relation *explanation** qui ne correspond à aucun exemple dans ANNODIS expliquant peut-être le manque de généralisation entre les deux types de données. Les prédictions du modèle artificiel construit surtout sur cette méta-relation pourraient être cohérentes expliquant l'amélioration observée. Les différences de performance au niveau des labels peuvent venir de distribution plus ou moins proche entre les deux types de donnée. Si on regarde la distribution en terme de traits (850 traits en tout), on constate un écart de plus de 30% pour 2 et 5 traits pour *result* et *explanation* mais aucun pour *contrast* et *continuation* pour lesquelles l'apport direct des données artificielles est positif.

5.3 Modèles avec sélection automatique d'exemples

Les expériences précédentes ont montré que l'ajout de données artificielles donnaient 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 manuelles.

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 on conserve les exemples prédits avec une probabilité supérieure à un seuil $s \in [30, 40, 50, 55, 60, 65, 70, 75]$. 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 une erreur de segmentation. On vérifie en quelque sorte aussi l'hypothèse de redondance du connecteur. Pour chaque seuil, on rééquilibre les données en se basant sur la relation la moins représentée (système+SELEC). A partir du seuil 80, ces expériences ne sont plus pertinentes, on conserve moins de 10 exemples par relation. Les seuils les plus intéressants sont les seuils 60, 65, 70 et 75 qui représentent respectivement un ajout de 553, 205, 72 et 16 exemples par relation. Les scores

des systèmes présentant les résultats les plus pertinents sont repris dans la table 5.

	MANONLY	ARTIONLY		UNION			MANKMAN			ADDPRED		ADDPROBA	INIT	INTERPLIN		
Seuil	-	60	70	60	70	75	30	65		40	65	65	65	60	70	75
Paramètre	-	-	-	-	-	-	250	0.5	900	-	-	-	-	0.7	0.9	0.7
Exactitude	39.7	27.0	23.8	26.2	38.1	41.7	35.3	30.6	**45.6**	42.5	44.4	**44.0**	43.3	36.5	31.7	34.9
contrast	13.3	32.0	29.5	26.7	26.8	11.6	16.4	**37.2**	32.0	14.5	31.6	24.7	24.0	34.1	30.7	24.6
result	49.0	20.0	8.2	25.4	41.7	50.0	29.5	27.8	**53.2**	47.4	52.6	**53.2**	47.8	33.6	15.2	29.7
continuation	39.7	8.6	16.5	19.4	39.1	43.3	**49.1**	20.6	38.5	36.8	40.6	43.4	43.4	28.8	19.8	27.0
explanation	43.8	31.8	32.1	30.9	39.7	45.6	35.3	34.3	51.1	**55.9**	45.9	44.4	48.9	46.1	51.0	52.9

TABLE 5 – Modèles avec sélection d'exemples, exactitude du système et f-mesure par relation

La sélection automatique d'exemples permet d'améliorer les résultats précédents, qu'il s'agisse du modèle ARTIONLY ou des modèles avec combinaison des données. De 23.0 d'exactitude avec ARTIONLY, on passe à 27.0 avec ARTIONLY+SELEC au seuil 60. De même on passe de 24.2 avec UNION à 41.7 avec UNION+SELEC au seuil 75, l'exactitude augmentant avec la croissance du seuil.

Il semble que les meilleurs systèmes soient obtenus entre les seuils 60 et 70. Au seuil 65, les systèmes INIT+SELEC, ADDPRED+SELEC et ADDPROBA+SELEC atteignent leur meilleur score (voir table 5), ce dernier améliorant significativement MANONLY (p-valeur= 0.046). L'exactitude de ces systèmes ne suit pas une évolution claire suivant le seuil. De même, si on retrouve avec INTERPLIN+SELEC une baisse de l'exactitude suivant α à chaque seuil, on n'a pas d'influence des seuils sur l'exactitude aux valeurs extrêmes de α.

Avec ARTIKMAN+SELEC et MANKMAN+SELEC on a la même tendance qu'avant, l'exactitude respectivement décroît et croît avec la croissance du coefficient pour chaque seuil, mais pour ARTIKMAN+SELEC on n'a rapidement plus assez d'exemples artificiels pour extraire des sous-ensembles. Pour MANKMAN+SELEC, l'exactitude avec le coefficient le plus bas augmente avec le seuil, de 22.6 (seuil 30) à 37.7 (seuil 75). C'est avec ce système et une influence très faible des données artificielles qu'on obtient le meilleur score d'exactitude, 45.6 améliorant significativement les performances de MANONLY (p-valeur= 0.021).

Au niveau des scores par relation, de nouveau une influence forte des données artificielles améliore l'identification de *contrast* avec en plus une influence positive d'un seuil haut mais inférieur à 70, au-delà le nombre d'exemples artificiels étant probablement trop bas pour influer sur l'identification (voir table 5). Parallèlement, à part avec ARTIONLY+SELEC, l'identification des autres relations s'améliore avec la croissance du seuil donc une baisse de l'influence du modèle artificiel. Pour *continuation* on observe toujours une amélioration pour une influence similaire des deux types de données et pour *explanation*, c'est toujours l'ajout de traits de prédictions qui permet les meilleures performances. Il semble qu'en plus on améliore ici l'identification de *result* (MANKMAN+SELEC et ADDPROBA+SELEC, table 5).

La sélection des exemples améliore l'identification des relations et conduit à deux systèmes améliorant significativement l'exactitude de MANONLY montrant que les données artificielles lorsqu'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.

A la constitution des corpus avec sélection on observe qu'avec la croissance du seuil on conserve toujours plus d'exemples pour *result*, dès le seuil 40 environ 3900 de plus, alors que *contrast* 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 celle dont les résultats sont les plus dégradés est celle pour laquelle il est le plus confiant.

6 Conclusion

Nous avons développé le premier système 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. Mais ces nouvelles données ne généralisent pas bien aux données implicites car elles sont de distribution différente. 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. Elles nous permettent des améliorations significatives par rapport au modèle n'utilisant que les données manuelles. Les meilleurs systèmes utilisent la sélection d'exemples et la pondération des données manuelles ou l'ajout de traits de prédictions du modèle artificiel.

Si 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. De plus, une étude des données explicites pourrait permettre d'augmenter la taille du corpus artificiel et d'améliorer sa qualité en sélectionnant des connecteurs et en identifiant des relations pour lesquelles cette méthode est plus ou moins efficace et des traits plus informatifs dans une optique de combinaison des données. Il faudra enfin porter ces méthodes sur les données anglaises pour une comparaison avec d'autres études.