Vers la détection automatique des affirmations inappropriées dans les articles scientifiques

Anna Koroleva[1]
(1) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
koroleva@limsi.fr

RESUME

Dans cet article nous considérons l'apport du Traitement Automatique des Langues (TAL) au problème de la détection automatique de « l'embellissement » (en anglais « spin ») des résultats de recherche dans les publications scientifiques du domaine biomédical. Nous cherchons à identifier les affirmations inappropriées dans les articles, c'est-à-dire les affirmations où l'effet positif du traitement étudié est plus grand que celui effectivement prouvé par la recherche. Après une description du problème de point de vue du TAL, nous présentons les pistes de recherche qui nous semblent les plus prometteuses pour automatiser la détection de l'embellissement. Ensuite nous analysons l'état de l'art sur les tâches comparables et présentons les premiers résultats obtenus dans notre projet avec des méthodes de base (grammaires locales) pour la tâche de l'extraction des entités spécifiques à notre objectif.

ABSTRACT

Towards automatic detection of inadequate claims in scientific articles
In this article we consider application of Natural Language Processing (NLP) techniques to the task of automatic detection of misrepresentation (« spin ») of research results in scientific publications from the biomedical domain. Our objective is to identify inadequate claims in medical articles, i.e. claims that state the beneficial effect of the experimental treatment to be greater than it is actually proven by the research results. After analyzing the problem from the point of view of NLP, we present methods that we consider applicable for automatic spin identification. We analyze the state of the art in similar or related tasks and we present our first results obtained with basic methods (local grammars) for the task of recognising entities specific for our goal.

MOTS-CLES : affirmations inappropriées ; spin ; embellissement de résultat, articles biomédicaux ; rôles sémantiques ; extraction des entités
KEYWORDS : inadequate reporting ; spin ; biomedical articles ; semantic roles ; entity extraction

1 Introduction

L'interprétation des résultats de recherche scientifique dans le domaine biomédical est souvent affectée par la présence d'embellissements ou d'affirmations inappropriées. Embellir les résultats d'un essai consiste à affirmer un effet positif (efficacité ou sûreté) du traitement étudié plus grand que celui prouvé dans l'essai (cf. Table 1) (Boutron et al., 2010 ; Boutron et al., 2014 ; Haneef et

al., 2015 ; Yavchitz et al., 2016). Le phénomène des affirmations inappropriées est appelé « spin » dans la littérature anglaise sur le sujet, et nous allons utiliser ce nom dans ce qui suit[1].

Résumé avec affirmation inappropriée	Résumé réécrit
*Treatment A **may be useful** in controlling cancer-related fatigue **in patients who present with severe fatigue** but is not useful in patients with mild or moderate fatigue.*	*Treatment A **was not more effective** than placebo in controlling cancer-related fatigue.*
*This study confirmed a previous trial demonstrating **improved PFS and response** for the treatment A compared with comparator B alone, although this did not result in improved survival.*	*The treatment A **was not more effective** than comparator B on overall survival in patients with metastatic breast cancer previously treated with anthracycline and taxanes.*

Table 1 : Exemples des résumés avec des affirmations inappropriées et des résumés réécrits par les experts afin que les conclusions correspondent aux résultats de la recherche. Les exemples ont été fournis par Isabelle Boutron et font partie de l'appendice de l'article (Boutron et al., 2014).

Plusieurs types d'essais médicaux se prêtent particulièrement bien à l'utilisation de spin, comme par exemple, les essais randomisés contrôlés[2] ou essais non-randomisés ainsi que les études d'exactitude de diagnostique (Boutron et al., 2010 ; Boutron et al., 2014 ; Lazarus et al., 2015 ; Yavchitz et al., 2016). En particulier, le spin est estimé être présent à différents degrés dans 60% des essais randomisés contrôlés qui sont la source la plus importante d'information de la médecine fondée sur les faits - *Evidence-Based Medicine* - (Boutron et al., 2014).

Le spin peut se trouver dans le résumé ainsi que dans le texte principal d'un article, mais il est plus fréquent dans les résumés (Boutron et al., 2010). Dans (Boutron et al., 2014), les auteurs ont étudié des articles du domaine de l'oncologie et ont prouvé que la présence de spin dans le résumé a de l'influence sur la manière dont les médecins interprètent l'effet positif du traitement discuté. Plus précisément, les médecins qui ont lu un résumé contenant du spin accordent plus d'efficacité au traitement que celle justifié par le corps de l'article. Le spin induit aussi une présentation inappropriée des résultats de recherche dans les communiqués de presse et les actualités de santé (Yavchitz et al., 2012 ; Haneef et al., 2015). Parfois le résumé est la seule partie disponible d'un article, augmentant ainsi fortement l'impact du spin sur la diffusion des résultats de recherche. La présence de spin influence les décisions prises par les médecins concernant l'utilisation des médicaments, le spin pose donc un problème grave, d'autant plus qu'il peut aussi affecter la perception que les patients ont d'un traitement. Il y a donc nécessité d'assister les auteurs d'articles scientifiques et leurs relecteurs à identifier les occurrences probables de spin. L'hypothèse explorée par notre projet est que les méthodes du Traitement Automatique des Langues (TAL) peuvent être utilisées pour détecter le spin en identifiant les affirmations inappropriées. Notre objectif est de développer des méthodes d'extraction des affirmations importantes et des éléments de preuve associés. Sur la base de la correspondance entre les affirmations et leurs informations justificatives nous allons calculer des traits caractéristiques qui seront utilisés par l'apprentissage automatique pour l'identification des textes contenant du spin. Dans ce but, nous allons collecter un corpus de textes biomédicaux et annoter les affirmations importantes, l'information justificative et l'absence/présence de spin.

[1] En français le terme « spin » est aussi utilisé, moins fréquemment on parle d'étude « embellie » ou encore de traitement « beautifié » cf. http://www.h2mw.eu/redactionmedicale/spin/

[2] Un essai randomisé contrôlé est un type d'étude scientifique utilisé pour tester l'efficacité d'un traitement. Les patients éligibles sont répartis au hasard en plusieurs groupes semblables : un (des) groupe(s) recevant le(s) traitement(s) étudié(s) et un groupe recevant un traitement de comparaison.

Dans cet article nous présentons la phase initiale du projet. Dans la section 2 nous présentons la notion de spin plus en détail, nous définissons les caractéristiques linguistiques des textes contenant du spin et examinons les fonctions de TAL que nous considérons comme nécessaires à son identification automatique. Dans la section 3 nous présentons les recherches en TAL qui ont abordé des tâches comparables à l'identification du spin et montrons comment nous comptons déployer les approches du TAL dans notre futur détecteur de spin. Dans la section 4 nous présentons nos recherches actuelles pour la tâche d'extraction des entités. Pour finir, dans la section 5 une conclusion et quelques perspectives seront présentées.

2 Notion de spin

Les études sur le spin ont distingué plusieurs types de spin, dont la fréquence varie en fonction du type d'essai médical abordé. Nous nous concentrons ici sur le spin dans les essais randomisés contrôlés (« *randomized controlled trials* », ou « *RCTs* »). Cette décision est basée sur l'importance de ce type d'essai par rapport aux autres types et est en accord avec les études existantes sur le spin. Dans ces dernières (Boutron et al., 2010 ; Lazarus et al., 2015 ; Yavchitz et al., 2016), les affirmations inappropriées sont classifiées en trois catégories et plusieurs sous-catégories :

1. présentation inappropriée des résultats de la recherche :
   - les effets négatifs ne sont pas présentés ;
   - on ne présente pas tous les résultats évalués (le résultat primaire n'est pas présenté ; on se concentre sur les résultats secondaires qui sont statistiquement significatifs)
   - la présentation du type et des caractéristiques de l'essai est imprécise ;
   - la présentation de la population étudiée est imprécise (on se concentre sur l'analyse de sous-groupes pour lesquels les résultats ont une significativité statistique ; on présente les résultats pour la population modifiée, par exemple après exclusion d'une partie des patients)
   - spin linguistique (utilisation des mots évaluatifs positifs de haut degré, comme *excellent*) ;
   - les limitations de l'essai ne sont pas présentées ;
   - les études précédentes sont partiellement citées (des articles importants ne sont pas cités) ;
2. interprétation inappropriée des résultats de la recherche :
   - on affirme que le médicament étudié a un effet positif ou l'effet équivalent au traitement de référence, bien que les résultats n'aient pas de significativité statistique ;
   - on présente le traitement comme sûr, bien que les résultats concernant la sûreté n'ont pas de significativité statistique ;
   - on affirme que le médicament étudié a un effet positif sans avoir fait de tests comparatifs ;
   - on considère la significativité statistique au lieu de la pertinence clinique ;
   - on affirme qu'il y a un effet causal entre le traitement étudié et le résultat évalué, bien que l'essai ne soit pas randomisé ;
3. extrapolation inappropriée :
   - au lieu de la population, l'intervention ou le résultat évalué, on présente une population plus large, une intervention ou un résultat différent ;
   - les implications tirées des résultats sont inappropriées pour la pratique clinique (conseils d'utiliser le traitement)

Le phénomène de spin est compliqué et hétérogène. Nous nous concentrons dans ce projet sur les types de spin identifiables uniquement à partir du texte de l'article, sans recourir à d'autres sources, par exemple les protocoles des recherches. Sur la base des types de spin, nous avons compilé une liste des tâches issues du domaine de TAL qui feront partie de l'identification automatique du spin :

1. extraction de l'évaluation du traitement (positif/neutre/négatif) ;
2. classification des articles biomédicaux selon le type d'essai (à ce moment nous nous sommes le plus intéressés à la distinction entre les essais randomisés contrôlés et les autres) ;
3. analyse de la structure du texte (division entre les parties titre/résumé/corps de l'article) ;
4. extraction des entités :
- extraction des résultats étudiés, de préférence en distinguant le résultat primaire (*primary outcome*) des résultats secondaires (*secondary outcome*)[3] ;
- extraction de la population et des groupes de patients ;
- extraction de la significativité statistique des résultats ;
- extraction des restrictions de l'essai ;
- extraction des effets négatifs ;
- extraction du traitement étudié et du traitement de référence.

5. extraction des relations entre les entités (par exemple relation entre résultats et significativité statistique)
6. analyse des paraphrases :
- le résultat principal déclaré dans le corps de l'article est-il présenté dans le résumé ?
- la population présentée dans le résumé est-elle la même que la population étudiée déclarée dans le corps de l'article, ou est-elle plus large (extrapolation), ou plus petite (sous-groupe) ?

7. analyse syntaxique : identification des constructions spécifiques au spin, par exemple proposition concessive associée souvent avec un changement de focus :

*« Treatment A may be useful in controlling cancer-related fatigue in patients with severe fatigue **but** is not useful in patients with mild or moderate fatigue »* (focus sur sous-groupe de patients).

*« This study demonstrates improved PFS and response for the treatment A compared with comparator B, **although** this did not result in improved survival »* (focus sur résultats secondaires).

3 Etat des recherches

Dans cette section nous allons considérer l'état des recherches en TAL pour deux tâches : 1) l'évaluation du biais, qui est lié à la détection de spin ; 2) l'extraction des entités et des relations entres elles, qui sont nécessaires à l'identification du spin.

3.1 Évaluation du biais (*bias assessment*)

Dans la littérature biomédicale, la tâche la plus proche de l'identification du spin est l'évaluation du biais dans les études cliniques. Le biais est défini comme une erreur systématique ou une déviation par rapport à la réalité dans les résultats ou les conclusions, qui peut induire une sous-estimation ou une surestimation de l'effet du traitement examiné (Higgins and Green, eds, 2008). Les erreurs peuvent concerner la conception de l'étude (*study design*), le déroulement de la recherche, l'analyse et la présentation des résultats. Les types de biais adressés incluent :

- le biais de sélection (*selection bias*) : génération de la séquence aléatoire ; masquage de l'allocation du traitement;
- le biais de performance (*performance bias*) : anonymisation (ou insu[4)] des participants et du personnel ;

[3] Les résultats (*« outcomes »)* sont les variables mesurées dans un essai. Elles doivent être déterminées à l'avance et divisées en primaires (les plus importantes pour prouver l'effet favorable ou défavorable du traitement) et secondaires (utilisées pour estimer les effets supplémentaires).

– le biais de détection (*detection bias*) : anonymisation de l'évaluation des résultats ;
– le biais d'attrition (*attrition bias*) : données incomplètes sur les résultats ;
– le biais de présentation (*reporting bias*) : présentation sélective des résultats.

Le biais de présentation nous intéresse le plus parce qu'il tombe aussi sous la notion du spin.

Plusieurs outils existent pour l'évaluation du risque de biais. Selon (Higgins et al., 2011), l'évaluation est le plus souvent faite par des experts en utilisant les méthodes autre que le TAL : des échelles (*scales*) et des listes de contrôle (*checklists*). Il y a aussi des approches qui explorent les méthodes de TAL pour automatiser l'évaluation de biais (Marshall et al., 2015-1, Marshall et al., 2015-2). Dans (Marshall et al., 2015-1), les auteurs décrivent la création d'un corpus sur la base des revues systématiques archivées et mises à disposition par le réseau Cochrane (*Cochrane Database of Systematic Reviews*, *CDSR*[5]). Cochrane[6] est un réseau international indépendant de chercheurs, professionnels de santé et patients dont le but d'améliorer la prise de décision dans le domaine de la santé. Les revues systématiques du CDSR contiennent l'information structurée (y compris l'évaluation de biais) sur les études, et c'est cette information qui est utilisée comme annotation du corpus. Les auteurs remarquent que l'évaluation de biais est subjective ; l'accord entre les experts étant le plus bas pour le biais de présentation. Il faut noter aussi que l'évaluation du biais de présentation est à la base des protocoles d'études. Les auteurs utilisent le modèle de machine à vecteurs support (« *Support Vector Machine* ») pour la classification des articles selon le niveau de biais, en prenant les mots du texte comme caractéristiques (approche sac de mots). Le système est présenté plus en détails dans (Marshall et al., 2015-2) où les auteurs comparent différents modèles pour l'évaluation de biais. Les résultats montrent que les données pour les biais d'attrition et de présentation sont très bruitées. Le modèle proposé par les auteurs a de meilleures performances pour les autres types de biais lorsque les données des biais d'attrition et de présentation sont exclues.

En conclusion, la tâche d'évaluation de biais est liée à la détection de spin car les deux regardent le problème de la présentation sélective. Les méthodes de TAL ont été utilisées pour l'évaluation de biais. Les recherches ont montré que l'accord entre les experts dans ce domaine est bas est que les données associées à certaines catégories de biais sont très bruitées, ce qui est problématique pour l'apprentissage automatique. La différence entre les deux tâches est que l'évaluation de biais considère la correspondance entre le protocole de la recherche et la présentation des résultats, tandis que nous n'analysons pas les protocoles pour le moment. Nous considérons la correspondance entre les caractéristiques d'une étude décrite dans le corps d'un article et celles présentées dans le résumé.

3.2 Extraction des entités et des relations

Le volume de recherche en extraction des entités de textes biomédicaux est assez grand, mais la plupart de la recherche est concentrée sur l'extraction des noms des gènes, protéines et médicaments. L'extraction des entités importantes pour nous a moins attiré l'attention (Summerscales et al., 2011).

---

[4] L'anonymisation (ou insu, par exemple on parle de « levée d'insu ») est utilisé dans les essais médicaux comparant plusieurs médicaments (des traitements expérimentaux et des traitements de référence). On s'assure que les patients ou les examinateurs, ou les deux, ne savent pas qui reçoit quel traitement. Comme pour l'évaluation d'articles scientifique ou parle alors de tests en « simple aveugle », « double aveugle » ou moins fréquemment de tests en « simple insu » ou « double insu » (http://www.spc.univ-lyon1.fr/polycop/double%20aveugle.htm).

[5] Cochrane Database of Systematic Reviews (CDSR) est une ressource primaire pour les revues systématiques dans le domaine médical. CDSR inclut des revues systématiques (les travaux dont le but est la synthèse des connaissances sur un sujet donné), les protocoles des revues et les éditoriaux.

[6] http://www.cochrane.org

Quand on parle des entités ici, on sous-entend les entités nommées (les noms propres comme les noms des médicaments ou des maladies), mais aussi les phrases (nominales, verbales ou autres) qui remplissent des « slots » (rôles sémantiques) dans le cadre sémantique décrivant les essais randomisés contrôlés : les résultats examinés, la population de patients, les effets négatifs secondaires, etc. (en s'inspirant de travaux comme Kiritchenko et al.,2010, Nguyen et al., 2013).

Cependant, la tâche d'extraction des entités concernant les caractéristiques des essais ou les descriptions des situations cliniques (y compris les informations concernant les patients, les traitements, les maladies ; les paramètres de l'essai, comme les résultats en question, les branches de l'intervention[7], les effets négatifs de l'intervention, etc.) a attiré de l'attention car elle est requise pour de nombreuses applications : résumé automatisé des textes, construction des revues systématiques, systèmes d'aide à la décision clinique, systèmes de réponse à des questions, création des bases de données structurées, recherche dans les bases de données, etc. Les approches utilisées pour l'extraction des entités varient en fonction du domaine et de la tâche spécifique abordée.

Pour la construction des revues systématiques, il faut trouver dans les articles sur les essais randomisés contrôlés au moins quatre éléments de base pour les études cliniques : population/problème, traitement, traitement de comparaison, et résultat (ces 4 éléments sont connus sous l'acronyme «*PICO* framework»[8] pour « *Population/Problem, Intervention, Comparator, Outcome* »). Mais il ne s'agit pas toujours de l'extraction des entités, car il suffit de trouver des phrases qui contiennent l'information recherchée sous une autre forme (Wallace et al., 2016).

(Dawes et al., 2007) regardent un ensemble plus large d'éléments : « *Patient–Population–Problem, Exposure–Intervention, Comparison, Outcome, Duration and Results* (*PECODR*) ». Les auteurs se concentrent sur les résumés. La tâche de cette recherche n'est pas encore de créer un système d'extraction des entités, mais d'explorer la possibilité d'extraire automatiquement ces types d'entités. Dans ce but, les auteurs retrouvent les termes qui sont associés à chaque élément (les contextes possibles comme « *outcome* » ou « *end point* » pour le résultat (*Outcome*), et aussi les mots qui réfèrent directement à l'entité, comme « *mortality* » pour l'entité *Outcome*). Le nombre de résumés étudiés est très petit (20 résumés). Sur la base de ce corpus, ils font la conclusion qu'il y a des mots spécifiques associés avec certains éléments (« *Comparison* », « *Outcome* » et « *Results* ») qui sont utilisés dans plusieurs articles, alors que les autres éléments (« *Patient–Problem* » et « *Exposure* ») ne sont pas définis par un ensemble de mots communs pour un grand nombre d'articles.

(De Bruijn et al., 2008) élargissent encore l'ensemble d'éléments, leur objectif est d'extraire des textes des essais randomisés contrôlés tous les éléments du *«Standards fusionnés dans la rédaction d'essais thérapeutiques» («CONSORT statement»*[9]) comme : les critères d'éligibilité, le nom du traitement étudié et le traitement de comparaison, les paramètres de l'intervention, y compris le dosage, la fréquence, la durée, etc., la taille de l'échantillon, la date de début et de fin d'inscription, les résultats primaires et secondaires avec les indicateurs temporels associés; l'information sur le financement, les métadonnées de la publication, y compris la date, les auteurs. L'objectif général est

[7] Les descriptions des branches de l'intervention comprennent normalement le nom du traitement expérimental et du traitement de référence, elles peuvent aussi inclure le dosage, les horaires et autres détails concernant l'administration.

[8] https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0029906/

[9] Le « Consolidated Standards of Reporting Trials » ou « CONSORT group » est un groupe regroupant des experts en méthodologie d'essais cliniques, experts en bonnes pratiques cliniques, des éditeurs de journaux du domaine biomédical et de sponsors de la recherche biomédicale dont le but est d'améliorer la qualité de la rédaction des essais randomisés contrôlés. Le «CONSORT Statement» est une liste d'items à vérifier pour évaluer la rédaction d'un essai. Site officiel: http://www.consort-statement.org/

d'extraire l'information essentielle sur les études qui puisse être utilisée dans des bases de données structurées. L'algorithme procède en deux étapes. D'abord, on utilise un classificateur pour choisir 5 phrases les plus pertinentes pour chaque élément. Ensuite, on applique des règles pour l'extraction des entités des phrases choisies. Les auteurs ont choisi comme mesure d'évaluation la correspondance partielle entre les séquences des mots obtenues et l'entité ciblée. La coïncidence totale entre les séquences trouvées et les entités n'est pas évaluée car déterminer les limites précises des entités est une tâche difficile. Pour certains types d'information (auteurs, DOI, date de publication), ce sont les métadonnées de la base de citation et de résumés d'articles du domaine biomédical Medline qui sont utilisées.

Ce système se distingue de la majorité des autres approches par le fait qu'il travaille avec des articles entiers au lieu des seuls résumés. L'approche a été développée dans les recherches suivantes des mêmes auteurs (Kiritchenko et al., 2010) qui ont créé le système appelé ExaCT pour l'extraction des entités des textes des essais randomisés contrôlés. Le système a pour objectif d'aider les examinateurs à identifier l'information nécessaire mais n'a pas pour objectif l'automatisation complète de l'extraction exhaustive de l'information.

Plusieurs études concernent l'extraction des entités pour le résumé automatisé de textes (Summerscales et al., 2009 ; Summerscales et al., 2011). Dans le premier les auteurs présentent les résultats préliminaires d'application des méthodes d'extraction des entités aux résumés des articles sur les études contrôlées du BMJ[10]. L'objectif est de trouver les noms des traitements, les groupes expérimentaux et les résultats examinés. Les auteurs décrivent quelques difficultés liées à ces types d'entités, telles que l'hétérogénéité des entités ciblées (une entité peut être représentée par un nom simple, par une phrase complexe, par une ellipse ; les traitements et résultats manquent de caractéristiques orthographiques spécifiques comme des chiffres, des lettres majuscules, ou des caractères spéciaux). Les auteurs utilisent le modèle de champs aléatoires conditionnels (« C*onditional Random Fields » ou «CRF»*) pour la classification des entités. Ils font des comparaisons entre différents ensembles de caractéristiques utilisées pour la classification: le mot ; sa catégorie lexicale ; l'identifiant de thématique médicale (« *Medical Subject Heading Id »*) associé au mot ; des étiquettes sémantiques (ici les types sémantiques des mots définis dans le métathesaurus UMLS[11], qui sont trouvés avec l'aide du programme MetaMap (Aronson, 2001) ; le titre de la section du résumé où le mot apparaît ; les quadrigrammes gauche et droite avec leurs catégories lexicales et leurs étiquettes sémantiques. Les auteurs font la conclusion que les caractéristiques qui comprennent les catégories thématiques et sémantiques du mot sont plus utiles que celles basées sur sa forme, notamment pour l'extraction des résultats examinés. Les auteurs supposent que l'utilisation des caractéristiques syntaxiques pourrait améliorer la performance de l'algorithme.

Dans (Summerscales et al., 2011) les auteurs présentent un système pour le calcul automatique de statistiques sommaires pour les essais randomisés contrôlés, y compris des valeurs de réduction de risque absolu (*Absolute Risk Reduction*, *ARR*)[12] et du nombre de sujets à traiter[13]. A cette fin, il est

[10] Le « British Medical Journal » (appelé officiellement « BMJ » depuis 1988 et « The BMJ » depuis 2014) est une revue britannique de médecine générale, l'une des revues médicales les plus anciennes et les plus lues dans le monde. Site officiel: http://www.bmj.com/

[11] L'UMLS (Unified Medical Language System) est un compendium des plusieurs vocabulaires contrôlés médicaux. Site officiel: https://www.nlm.nih.gov/research/umls/

[12] La réduction du risque absolu est la différence entre la probabilité de survenue d'un événement dans le groupe expérimental et la probabilité de survenue de cet événement dans le groupe témoin (groupe qui reçoit le traitement de référence).

[13] Le nombre de sujets à traiter est le nombre de patients qu'il est nécessaire de traiter pendant une période donnée pour éviter 1 événement défavorable.

nécessaire d'extraire les descriptions et les tailles des groupes de patients, les résultats étudiés et le nombre de résultats positifs/négatifs. D'abord l'algorithme cherche à identifier les phrases contenant l'information pertinente. Ensuite, un classificateur à base de CRF est appliqué pour trouver les mentions des groupes expérimentaux, des résultats examinés, la taille des groupes expérimentaux et le nombre de résultats positifs/négatifs. L'algorithme cherche ensuite à établir des relations entre les entités (taille du groupe + définition du groupe ; résultat étudié + nombre de résultats positifs/négatifs) et à calculer des statistiques sommaires sur la base des informations extraites. Cette recherche montre que les résultats examinés (« outcomes ») sont plus difficiles à extraire que les autres types d'informations.

(Chung, 2009) se concentre elle, sur l'extraction des « branches d'intervention », c-a-d les différents traitements ou placebos mis en jeux lors de l'essai, souvent associés chacun à un groupe de patients, des essais randomisés contrôlés. L'auteur explore la possibilité d'extraire les branches de l'intervention sur la base de l'analyse des constructions coordonnées. L'approche est basée sur l'analyse des phrases de la section « Méthodes » des résumés où l'information concernant les branches est normalement présentées explicitement ; les sections « Objectifs », « Résultats », et « Conclusions », où l'information pertinente peut apparaître implicitement, ne sont pas concernées. L'auteur a collecté et analysé un corpus d'essais randomisés contrôlés et a abouti à la conclusion que les constructions coordonnées sont fréquentes dans les descriptions de branches. L'auteur explore l'hypothèse que l'analyse syntaxique automatique des phrases complètes puisse être utilisée pour trouver les constructions coordonnées qui décrivent les branches de l'intervention. L'algorithme comprend deux étapes, d''abord les phrases du résumé sont classées pour identifier celles qui appartiennent à la section méthodes. Ensuite, les phrases sont normalisées, et un classifieur basé sur l'entropie maximale est utilisé avec des ensembles de caractéristiques différents (conjonction de coordination ; type de coordination ; arguments syntaxiques, etc.). Les classes sémantiques des mots sont affectées sur la base de l'UMLS avec l'aide de MetaMap.

D'autres recherches se sont intéressées à l'extraction des informations concernant la population des patients. Par exemple, (Xu et al., 2007) proposent une approche pour extraire les descriptions de population, le nombre de patients examinés et les descriptions des maladies/symptômes des essais randomisés contrôlés. L'algorithme procède en deux étapes. La première consiste à rechercher des phrases qui puissent contenir l'information sur la population ; un modèle de Markov caché est utilisé. La deuxième étape concerne l'extraction des entités avec des règles syntaxiques. Une approche similaire est explorée dans (Raja et al., 2016), les auteurs utilisent un classifieur binaire pour trouver les phrases contenant l'information sur la population ; ensuite ils appliquent des règles utilisant des expressions régulières et une analyse syntaxique pour extraire les entités.

De ces travaux antérieurs, nous pouvons tirer quelques conclusions :

- la plupart des recherches sont concentrées sur les essais randomisés contrôlés ;
- la plupart des recherches considèrent l'extraction des entités seulement à partir des résumés ;
- l'approche commune comporte deux étapes : filtrer les phrases qui puissent contenir les entités ; trouver les entités dans les phrases filtrées.
- la plupart des approches combinent l'apprentissage automatique et les règles ;
- Des systèmes comme MetaMap pour associer les termes utilisés dans les articles aux termes de l'UMLS sont utilisés souvent ;
- définir les limites précises d'une entité est une tâche plus difficile qu'obtenir une correspondance partielle ;
- le résultat examiné (*outcome*) est une des entités les plus difficiles à identifier.

4 Expériences

4.1 Le corpus

Nous allons collecter un corpus pour la détection du spin et l'annoter avec l'information pertinente plus tard au cours de notre projet. En ce moment, nous utilisons pour l'analyse préliminaire un corpus d'articles de PMC[14] collecté lors d'une expérience précédente effectuée au laboratoire. Nous avons identifié 3938 articles concernant des essais randomisés contrôlés dans notre corpus sur un total de 65396 articles.

4.2 Extraction des entités nommées : résultat examiné (*outcome*)

L'entité la plus importante pour l'identification de spin est le résultat examiné (*outcome*) parce que plusieurs types de spin concernent ce type d'entité. Nous allons utiliser le mot anglais *outcome* dans cette section pour ne pas confondre le nom de l'entité et le mot *résultat* avec son sens usuel.

Nous avons mentionné plusieurs études qui abordent la tâche d'extraction de l' « outcome ». Ces études considèrent les phrases comme l'exemple suivant qui provient de l'article (Summerscales et al., 2009) : «*Mortality was higher in the quinine than in the artemether*». Les auteurs remarquent que sauf pendant la présentation des résultats, l' « outcome » est normalement mentionné explicitement dans le résumé ou dans le corps de l'article ou bien dans les deux, ceci avec des expressions du type :

- *Expected outcome is… (« le résultat attendu est ... »)*
- *... was our primary outcome. (« ... était notre résultat principal »)*
- *Outcomes include...(« le résultat inclut ... »)*

Les études citées ci-dessus ne cherchent pas à identifier ce genre d'expressions car leur objectif est d'extraire les résultats obtenus et présentés par les recherches. Notre but est différent : retrouver les descriptions des « outcomes » initiaux d'un essai ainsi que les « outcomes » présentés dans l'article étudié, pour identifier les cas où ils ne correspondraient pas. Nous pouvons donc nous concentrer en première approche sur les constructions données en exemple précédemment (« *outcome is ...* »), ce qui est une tâche plus simple que celles présentées dans l'état de l'art de la section 3.2.

Parmi les types de spin les plus fréquents, nous trouvons l'absence de l' « outcome » primaire dans la présentation des résultats et l'accent mis sur les « outcomes » secondaires. Nous nous sommes donc intéressés à la distinction entre les deux types d' « outcome » : primaire et secondaire. Actuellement notre but est d'évaluer la possibilité d'extraire les « outcomes » des descriptions explicites et d'identifier l' « outcome » primaire. Notre recherche est basée sur 3938 textes de notre corpus qui ont le type «*Randomized controlled trial* ». Nous procédons selon les étapes suivantes :

1. Nous avons identifié manuellement toutes les phrases du corpus qui contiennent des mots qui puissent être utilisés pour décrire des outcomes :
   - les mots « *outcome* », « *endpoint* » (avec ses variantes orthographiques « *end point* », « *end-point* ») qui décrivent directement l'entité « outcome » ;
   - les noms « *aim* », « *purpose* », « *objective* », etc., les verbes « *explore* », « *investigate* », « *examine* », etc., qui qui peuvent servir comme marqueurs de l'entité « *outcome* ».

[14] PMC (PubMed Central) est une base de données d'articles en génie biomédical et dans les sciences de la vie. Site officiel : https://www.ncbi.nlm.nih.gov/pmc/

2. Nous avons analysé les phrases ainsi obtenues pour trouver les constructions fréquentes utilisées pour décrire l' « outcome ». Nous avons porté une attention spécifique à deux points importants : a) la distinction entre la description la plus explicite des « outcomes » (avec les mots « *outcome* » et « *endpoint* ») et les autres variantes des descriptions de l'objectif ; b) la distinction entre l' « outcome » primaire et les « outcome » secondaires.
3. Nous avons créé des règles à base de grammaires locales avec la boîte à outils Unitex (Paumier, 2016) pour trouver ces constructions. La figure 1 donne un exemple de graphe pour identifier les « outcomes ». Actuellement nous utilisons 9 graphes pour détecter tous les « outcomes » possibles.

Nous cherchons aussi à séparer le résumé du corps de texte. Dans ce but, nous avons créé des règles qui réussissent à trouver les résumés pour 3935 articles de notre corpus.

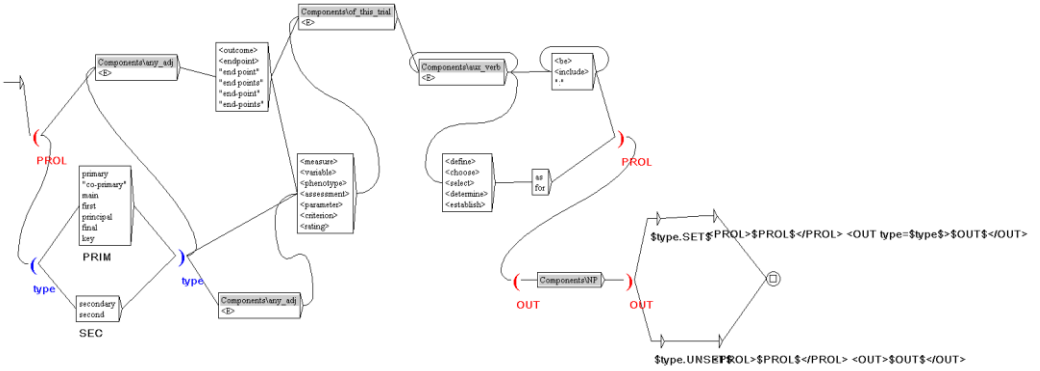

Figure 1 : Un exemple du graphe pour identifier les « outcomes »

Il y a plusieurs raisons pour lesquelles nous utilisons des règles pour la détection des parties des articles et pour l'extraction des entités. L'approche basée sur des règles est une alternative à l'apprentissage automatique pour ces tâches. Par exemple, CasSys et CasEn sont deux systèmes pour l'extraction des entités employant des règles à base de grammaires locales (Friburger et al., 2004 ; Maurel et al., 2011). L'autre raison pour employer les règles au stade actuel de notre projet est l'absence de corpus annoté pour notre recherche. La collecte d'un corpus des textes contenant du spin fait partie des étapes suivantes de notre travail. Dès que le corpus sera annoté avec l'information nécessaire pour identifier le spin, y compris les entités, nous pourrons employer des méthodes d'apprentissage automatique et comparer leur performance avec celle du système basé sur des règles. Le système actuel servira donc comme une approche de référence. En plus, nous comptons utiliser les grammaires locales pour une pré-annotation du corpus, avant annotation par les annotateurs humains. En absence de corpus annoté nous n'avons pas d'évaluations précises de la performance de l'algorithme actuel, les estimations préliminaires ont été effectuées sur un petit échantillon d'exemples compilés manuellement.

Un exemple des résultats que nous obtenons est montré dans la table 2, avec les notations suivantes :

- *Location* correspond à la partie du texte (*Abstr* - le résumé, *Body_text* - le texte principal).
- *Extracted words* sont les séquences des mots que nous avons identifiées comme candidates pour être un « outcome ».
- *Sentence* est une phrase où l'on a trouvé une occurrence d'une construction. <PROL> marque le contexte de l'entité candidate ; <OUT> marque l'entité.

Location	*Extracted words*	*Sentence*
Abstr	the remission of depressive symptoms at the 2-month follow up visit	The <PROL>primary outcome was</PROL> <OUT type=PRIM>the remission of depressive symptoms at the 2-month follow up visit</OUT>, defined as a HDRS score of 7 or less.
Body_text	overall mortality, severity of BPD, number of days on the ventilator, number of treatment failures, ventilation-induced lung injury and pulmonary hypertension	<PROL>Secondary outcome parameters are</PROL> <OUT type=SEC>overall mortality, severity of BPD, number of days on the ventilator, number of treatment failures, ventilation-induced lung injury and pulmonary hypertension</OUT> according to clinical and laboratory parameters and need for ECMO (only for ECMO centres).

Table 2 : Résultat d'application des règles

4.3 Extraction des entités nommées: population/groupes des patients

Pour la tâche d'extraction de population/groupes des patients notre objectif est identique aux recherches précédentes. Pour le moment nous ne regardons pas spécialement l'extraction du nombre de participants, mais seulement les descriptions de patients (âge, sexe, nationalité). Nous avons utilisé la même approche que pour l' « outcome »: nous avons collecté toutes les phrases qui contiennent une liste des mots pertinents (« *adults* », « *children* », « *adolescents* », « *population* », « *cohort* », etc.). Ensuite nous avons analysé les constructions qui décrivent la population étudiée. La figure 2 donne un exemple du graphe utilisé pour identifier la population étudiée. Actuellement nous utilisons 5 graphes pour détecter les descriptions de la population.

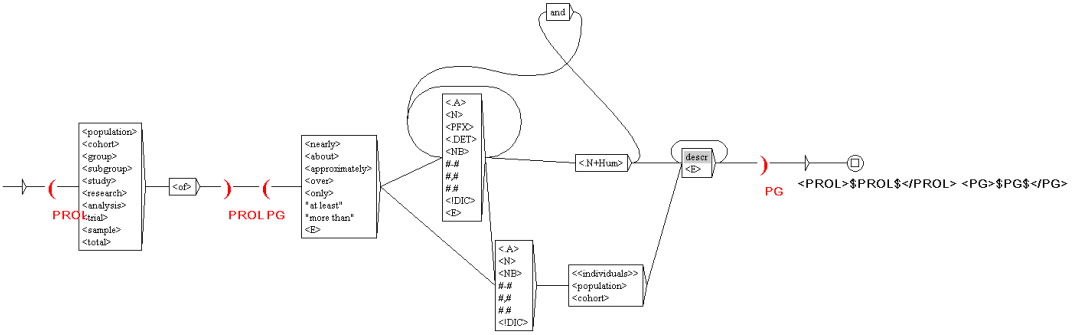

Figure 2 : Un exemple du graphe pour identifier la population étudiée

Pour faire la recherche la plus stricte possible, nous limitons les séquences des mots trouvées (notées par <...>) en exigeant qu'elles contiennent des mots du groupe sémantique « humains », que nous déterminons par une liste des mots compilée manuellement et par le trait <.N+HUM> des dictionnaires Unitex. Nous avons inclut aussi la construction plus générale qui contient des mots comme « *adults* », « *children* », « *adolescents* », etc. précédé ou suivis par une description (adjectif, nombre, groupe prépositionnel). Pour cette construction nous n'utilisons pas le trait <.N+HUM>.

Les constructions ci-dessus ont été trouvées dans 3935 textes (99,9%) de notre corpus. Mais la précision de la recherche est assez basse: notre analyse montre que les constructions étudiées sont souvent utilisées dans les rubriques des articles dédiées à la revue de la littérature, pour décrire la population examinée dans les études antérieures, ce que nous ne voulons pas extraire. Notre tâche future sera de créer des algorithmes pour filtrer les constructions trouvées.

5 Conclusion et perspectives

Dans ces premières expériences sur la détection du spin dans les articles scientifiques, nous avons examiné la possibilité d'extraire les résultats (« outcomes ») d'un essai randomisé contrôlé et la population ou les groupes des patients, avec l'aide de grammaires locales. Nous avons mis l'attention particulièrement sur le « primary outcome » (résultat primaire), qui est crucial pour la détection de spin, sur la base des constructions qui décrivent les objectifs et les résultats étudiés.

Nos résultats montrent que les constructions simples qui décrivent les « outcomes » et la population couvrent une grande partie de textes (99,9% pour la population, 91,5% et 94% pour les constructions les plus générales concernant l' « outcome » examiné). Un algorithme du filtrage automatique des résultats trouvés est cependant encore nécessaire pour identifier les entités ciblées.

Concernant les constructions plus spécifiques, une mention explicite du « primary outcome » est trouvée dans 51,8% d'articles, une mention de l'objectif principal (du type « *We aim at <...>* » / « *Our main goal is <...>* », etc.) est présente dans 21,1% des articles. La construction « *<...> is assessed/investigated/studied/analyzed/explored* » est utilisée pour décrire des outcomes, mais aussi les groupes de patients et toutes les mesures utilisées dans une étude. Elle est la plus fréquente (apparaît dans 94% des textes). Nous supposons qu'en l'absence d'une mention explicite d' « outcome », on peut identifier le « primary outcome » sur la base des mentions de l'objectif principal et des mesures utilisées. Cependant, cette hypothèse doit être discutée avec les experts et vérifiée sur un corpus annoté. Une analyse préliminaire montre qu'afin d'extraire l' « outcome » des constructions générales, il faudrait développer des algorithmes supplémentaires, tel que:

- des règles de filtrage pour éliminer les phrases des types « *Outcomes were measured* » et « *Study subjects were evaluated every two weeks* ». Dans la recherche future, nous allons tester des règles simples basées sur les mots indésirables (*outcome*, *variable*, etc.) et l'intersection avec l'entité *Patient group/Population* ;
- des règles d'extraction des mentions de l'objectif lorsque les mentions sont très générales (e.g. « *Our aim is **to evaluate efficacy*** » dont on doit extraire « efficacy »).

Nos prochains travaux concerneront l'extraction de l' « outcome » à partir de références implicites, l'association des « outcomes » avec leur significativité statistique, et la vérification de la correspondance entre les « outcomes » décrits dans le corps d'un article et dans les sections « Résultats » et « Conclusions ». Nous sommes intéressés par les phrases du type suivant : « ***Survival rate** was higher in group A **(p=0.0003)** ».* Etant donné la difficulté d'extraire l' « outcome » dans de telles phrases, nous allons explorer la possibilité d'extraire plusieurs phrases candidates. Nous comptons aussi investiguer l'emploi de traitement spécifique lorsque les rôles sémantiques sont remplis par des entités nommées et bien sûr approfondir notre modèle de rôles sémantiques par rapport aux modèles utilisés pour l'extraction d'évènements, les « frames », etc.

Remerciements

Ces travaux ont été effectués dans le cadre du projet MIROR financé par le programme de recherche et d'innovation de l'Union Européenne Horizon 2020 sous la référence « Marie Sklodowska-Curie grant agreement No 676207 ».

Références

ARONSON A. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. Proc. *AMIA Symposium*.

BOUTRON I., ALTMAN D.G., HOPEWELL S., VERA-BADILLO F., TANNOCK I., RAVAUD P. (2014). Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of Cancer: the SPIIN randomized controlled trial. *J Clin Oncol*, 32, 4120–4126.

BOUTRON I., DUTTON S., RAVAUD P., ALTMAN D.G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 303, 2058–2064.

CHUNG G.Y. (2009). Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *J Biomed Inform*, 42(5), 790-800.

DAWES M., PLUYE P., SHEA L., GRAD R., GREENBERG A., NIE J.-N. (2007). The identification of clinically important elements within medical journal abstracts: Patient-population-problem, exposure-intervention, comparison, outcome, duration and results (PECODR). *Informatics in Primary Care*, 15(1), 9–16.

DE BRUIJN B., CARINI S., KIRITCHENKO S., MARTIN J., SIM I. (2008). Automated information extraction of key trial design elements from clinical trial publications. Proceedings of *the AMIA Annual Symposium*, 141-145.HANEEF R., LAZARUS C., RAVAUD P., YAVCHITZ A., BOUTRON I. (2015). Interpretation of results of studies evaluating an intervention highlighted in Google Health News: a cross-sectional study of news. *PLoS ONE*, 10(10), doi:10.1371/journal.pone.0140889.

FRIBURGER N., MAUREL D. (2004). Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, 313, 94-104.

HIGGINS J.P., ALTMAN D.G., GOTZSCHE P.C., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343:d5928

HIGGINS J.P., GREEN S., eds. (2008). Cochrane handbook for systematic reviews of interventions. West Sussex : Wiley & Sons Ltd.

KIRITCHENKO S., DE BRUIJN B., CARINI S., MARTIN J., SIM I. (2010). ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 10: 56-10.1186/1472-6947-10-56.

LAZARUS C., HANEEF R., RAVAUD P., BOUTRON I. (2015). Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*, 15:85.

MARSHALL I.J., KUIPER J., WALLACE B.C. (2015-1). Automating risk of bias assessment for clinical trials. *IEEE journal of biomedical and health informatics*, 19(4), 1406-1412.

MARSHALL I.J., KUIPER J., WALLACE B.C. (2015-2). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, ocv044.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D. (2011). Cascades autour de la reconnaissance des entités nommées. *TAL* 52-1.

NGUYEN N., MIWA M., TSURUOKA Y., TOJO S. (2013). Open information extraction from biomedical literature using predicate-argument structure patterns. *Proceedings of the 5th International Symposium on Languages in Biology and Medicine,* 51–55.

PAUMIER S. (2016). Unitex 3.1 User Manual. http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf

RAJA K., DASOT N., TECH B., GOYAL P., JONNALAGADDA S.R. (2016). Towards evidence-based precision medicine: extracting population information from biomedical text using binary classifiers and syntactic patterns. *AMIA Jt Summits Transl Sci Proc*, 203-212.

SUMMERSCALES R.L., ARGAMON S., BAI S., HUPERFF J., SCHWARTZFF A. (2011). Automatic summarization of results from clinical trials. *The 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 372–377.

SUMMERSCALES R.L., ARGAMON S., HUPERT J., SCHWARTZ A. (2009). Identifying treatments, groups, and outcomes in medical abstracts. *The Sixth Midwest Computational Linguistics Colloquium* (MCLC 2009). 2009.

WALLACE B.C., KUIPER J., SHARMA A., ZHU M., MARSHALL I.J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132), 1–25.

XU R., GARTEN Y., SUPEKAR K.S., DAS A.K., ALTMAN R.B., GARBER A.M. (2007). Extracting subject demographic information from abstracts of randomized clinical trial reports. *Proceedings of the 12th World Congress on Health (Medical) Informatics*. Brisbane, Australia. Edited by: Kuhn K.A., Warren J.R., Leong T.Y. Amsterdam: IOS Press ; 2007:550-554.

YAVCHITZ A., BOUTRON I., BAFETA A., MARROUN I., CHARLES P., MANTZ J., et al. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*, 9:e1001308.

YAVCHITZ A., RAVAUD P., ALTMAN D.G., MOHER D., HROBJARTSSON A., LASSERSON T., BOUTRON I. (2016). A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of Clinical Epidemiology*, 75, 56-65.