Cascade de CRFs et SVM pour la détection de références bibliographiques diffuses dans les articles scientifiques

Anaïs Ollagnier[*,**] **— Sébastien Fournier**[*] **— Patrice Bellot**[*,**]

[*] *Aix-Marseille Université, CNRS, LSIS UMR 7296, 13397 Marseille, France*
[**] *Aix-Marseille Université, CNRS, CLEO OpenEdition UMS 3287, 13451 Marseille, France*
anais.ollagnier, sebastien.fournier, patrice.bellot@lsis.org

RÉSUMÉ. *Dans le contexte d'une bibliothèque d'articles scientifiques, les références bibliographiques sont une source majeure de liens. Parmi elles, certaines sont explicites comme les références que nous pouvons retrouver à la fin des articles ou des livres. Tandis que d'autres sont dispersées selon un degré de diffusion plus ou moins fort dans le corps du texte. Nous proposons de nous focaliser sur la détection de ce type de références que nous nommons références bibliographiques diffuses présentes dans des articles scientifiques. Nous prenons exemple du domaine des Sciences Humaines et Sociales qui est particulièrement riche de telles citations. Afin de pallier les difficultés inhérentes à la détection de ce type de références, nous introduisons une méthodologie qui consiste, d'une part, à identifier les paragraphes qui contiennent des références via un processus de classification supervisée et, d'autre part, dans l'application de CCRFs afin de détecter plus précisément les zones bibliographiques et d'annoter leurs contenus.*

ABSTRACT. *In the context of a library of scientific articles, bibliographic references are a major source of links. Among them, some are explicit references as we can find at the end of articles or books. While others are scattered according to an implicit degree more or less strong in the text. We propose to focus on the detection of this type of references that we call diffuse in scientific articles from the field of Humanities and Social Sciences. To overcome the difficulties inherent in the detection of such references, we present a method which comprises, firstly, to identify paragraphs that contain references via a classification process and, secondly, in applying CCRFs to more accurately detect the bibliographic entries and annotate their contents.*

MOTS-CLÉS : *CCRFs*$_1$, *Système de recommandation*$_2$, *Détection de référence bibliographique*$_3$.

KEYWORDS: *CCRFs*$_1$, *Recommender system*$_2$, *Bibliographical reference detection*$_3$.

1. Introduction

Par définition, une référence bibliographique regroupe l'ensemble des éléments de données nécessaires pour identifier un document ou une partie de document de tout type, sur tout support (livre, article, site web, etc.). L'utilisation des références bibliographiques est un élément crucial dans la création et la diffusion d'informations (Cronin, 1984). Le travail présenté dans cet article s'intègre dans le cadre de la réalisation d'un système de recommandation de lectures axé sur l'exploitation de liens entre les contenus. L'exploitation de liens entre les contenus est un élément crucial des approches de recommandations. Nous nous intéressons particulièrement à un aspect inhérent à ce type de système : la caractérisation du contenu du document. Dans les systèmes de recommandation de lectures, l'utilisation du contenu au sein des systèmes de recommandation fonctionnent par la caractérisation du contenu de l'information à filtrer. Les représentations des documents et des profils dans ce type de filtrage exploitent seulement les informations qui peuvent être dérivées de leur thème respectif. Autrement dit, la sélection de documents est fondée sur une comparaison des sujets abordés dans des documents relatifs à des thèmes d'intérêt pour l'utilisateur. Or, de nombreuses informations, telles que les références bibliographiques et les entités nommées dispersées au sein du document peuvent s'avérer utiles lors de la recommandation car elles peuvent servir de liaison entre des informations et des documents directement liés au contenu ciblé. C'est dans cet intérêt que le travail présenté dans cet article se focalise sur la détection des références bibliographiques et plus particulièrement des références bibliographiques diffuses.

Pour rappel, l'extraction d'informations bibliographiques consiste à identifier automatiquement chaque mot d'une référence bibliographique comme l'un des champs bibliographiques prédéfinis, tels que, l'auteur, le titre, la date, etc. L'extraction d'informations bibliographiques est souvent restreinte aux zones les plus formelles, à savoir, les références que nous pouvons trouver généralement à la fin des articles ou des livres. Les résultats dans ce domaine montrent par ailleurs des performances très satisfaisantes (Kim *et al.*, 2011 ; Anzaroot et McCallum, 2013). Cependant, les références bibliographiques peuvent également se trouver à différents niveaux du document tels que les notes de bas de pages ou dans le corps du texte. Ce type de références que nous nommons références diffuses possèdent de nombreuses particularités comme une structuration et une répartition de ses éléments dans le texte très variées. Ces deux aspects sont influencés par le degré d'implicite de la référence. En effet, les références bibliographiques diffuses peuvent être plus ou moins explicites, c'est à dire, que nous pouvons les trouver disséminées dans le texte ou sous la forme de citations qui vont être normées par des conventions selon le type de publication. Ces différents aspects rendent difficiles leur identification par des méthodes classiques d'analyse de références. Le degré d'implicite de la référence peut la rendre difficilement identifiable via de simples caractéristiques lexicales ou morphologiques souvent exploités dans l'annotation automatique de références bibliographiques.

Notre contribution s'intègre autour de la réalisation d'une méthodologie dédiée à la détection de ces références bibliographiques diffuses. Cette méthodologie consiste, d'une part, à identifier les paragraphes qui contiennent des références via un processus

de classification supervisée et, d'autre part, dans l'application de CCRFs afin de détecter plus précisément les zones bibliographiques et d'annoter leurs contenus. L'originalité de notre contribution est d'exploiter les références diffuses en vue d'établir des liens entre les documents dans le cadre de la réalisation d'un système de recommandation de lectures. De plus, dans notre travail, nous traitons avec des données de Sciences Humaines et Sociales (SHS) qui est un défi, car en plus de traiter un domaine multidisciplinaire très diversifié la présence de nombreuses variations structurelles des références bibliographiques rend l'homogénéisation des traitements difficiles.

Le document est structuré de la manière suivante. La section 2 présente un état de l'art des méthodes traditionnelles de détection et d'annotation de références bibliographique ainsi que des outils déjà existant. La section 3 se focalise sur la présentation des données que nous allons utiliser afin de valider notre approche. La section 4 présente notre approche et la section 5 expose les expérimentations et les résultats. Enfin, la section 6 conclut.

2. Travaux connexes

Depuis les années 90, l'analyse automatique de références suscite beaucoup d'attention pas seulement pour leur utilisation dans les processus d'indexation mais aussi pour améliorer les performances lors de la récupération et l'extraction d'information. CiteSeer[1] (Giles *et al.*, 1998) qui est un moteur de recherche publique ainsi qu'une bibliothèque numérique pour les articles scientifiques appartenant au domaine informatique, est l'un des premiers systèmes d'indexation automatique de citations. Depuis, plusieurs outils permettent l'analyse du réseau de citation de la littérature scientifique avec une vocation disciplinaire ou pluridisciplinaire (Bellis, 2009). Au niveau disciplinaire, des bases de données fournissent aux spécialistes du domaine via des systèmes autonomes d'analyse de citations, notamment ceux fournis par le *Chemical Abstract Service*[2] (CAS) pour la chimie et d'autres sciences connexes ; le *SAO/NASA Astrophysics Data System*[3] (ADS) pour l'astronomie ; *IEEE Xplore*[4] pour le domaine informatique. Au niveau multidisciplinaire, nous pouvons citer *Google Schoolar*[5] et *Scopus*[6].
Le problème de l'analyse des références bibliographiques est considéré comme un problème de labellisation de séquence dans lequel la référence est considérée comme une chaîne de caractère. Beaucoup de travaux se sont penchés sur l'annotation de séquences selon des classes prédéfinies, nous pouvons classer ces travaux en trois grandes approches : expressions régulières basées sur des heuristiques, algorithmes d'apprentissage et des systèmes à base de connaissances. La première approche fon-

1. http://citeseerx.ist.psu.edu/
2. https://www.cas.org/
3. http://ads.harvard.edu/
4. http://ieeexplore.ieee.org/Xplore/home.jsp
5. https://scholar.google.fr/
6. http://www.scopus.com/

dée sur des expressions régulières basées sur des heuristiques, comme le définissent (Huang *et al.*, 2004), consiste à extraire des séquences redondantes via l'utilisation d'algorithmes heuristiques. Ces séquences permettent ensuite d'établir des modèles pour l'analyse des séquences. L'utilisation de ces approches basées sur des heuristiques sont un bon compromis entre la qualité des solutions trouvées et la rapidité de la mise en oeuvre du procédé, toutefois, ils ne garantissent pas nécessairement des solutions optimales selon la complexité de la tâche à réaliser (Deconinck, 2010). La seconde approche fondée sur des algorithmes d'apprentissage emploie des algorithmes qui opèrent en construisant un modèle basé sur un corpus donné en entrée. Le modèle construit procède ensuite à des prédictions ou des décisions afin d'analyser les séquences. Bien que ces approches ont une bonne capacité d'adaptation et de bons résultats (Anzaroot et McCallum, 2013), elles ont des limitations et, en particulier, le corpus d'apprentissage est souvent hautement dépendant des données d'application. La troisième approche fondée sur des systèmes à base de connaissances utilise, dans la plus part des cas, les connaissances de domaine pour dériver une ontologie qui décrit les données d'intérêt. Cette connaissance va alors inclure des informations comme les relations, les caractéristiques lexicales et des mots clés contextuels (Rezaei et Muntz, 2013). Par l'analyse de l'ontologie, plusieurs règles et extracteurs peuvent être ainsi générés. Ces règles et extracteurs sont ensuite utilisés pour effectuer l'extraction de l'information. Cependant, il nécessite un expert du domaine pour maintenir la base de connaissances.

En dehors des techniques largement employées pour l'analyse des références, de nombreux outils ont été réalisés. La plupart de ces outils sont basés sur des algorithmes d'apprentissage, en particulier sur les CRFs. ParsCit[7] (Councill *et al.*, 2008) basé sur des CRFs fournit une boite à outil qui permet l'annotation générale de références dans les zones bibliographiques. Biblio Citation Parser[8] est développé simulatanément que ParsCit, et il est également conçu pour les références à la fin des articles. Freecite[9] est également inspiré par ParsCit, il utilise des CRFs avec l'implémentation de librairies CRF++. Grobid[10] (Lopez, 2009) permet l'extraction et l'analyse de références bibliographiques sur la base de CRFs. Un point intéressant est que cet outil peut enrichir l'annotation via l'utilisation de données externes. Bibpro[11] (Chen *et al.*, 2012) capture les propriétés structurelles et transforme des propriétés en un modèle de séquence. Son modèle est basé sur l'alignement de séquence, des algorithmes d'apprentissage et des systèmes à base de connaissances. Bilbo[12] (Kim *et al.*, 2012a) permet l'annotation des références présentes dans les zones bibliographiques ainsi que celles présentes dans les notes de bas de pages. Ses modèles sont basés sur l'utilisation de CRFs mais également de machine à vecteurs de support (*Support Vector Machine* (SVM)) pour la classification des notes de bas de pages. Nous constatons, suite à l'étude de ces dif-

7. `http://aye.comp.nus.edu.sg/parsCit/`
8. `http://paracite.eprints.org/developers/`
9. `http://freecite.library.brown.edu/welcome`
10. `https://github.com/grobid/grobid`
11. `https://github.com/ice91/BibPro`
12. `https://github.com/OpenEdition/bilbo`

férents outils, qu'aucun ne permet l'annotation des références disséminées au sein du corps du texte.

Dans le cadre de notre travail, nous avons choisi d'utiliser une approche basée sur des algorithmes d'apprentissage, et plus particulièrement fondée sur des CRFs, parce qu'ils constituent un très bon compromis entre performances et rapidité de mise en oeuvre. De plus la littérature démontre leur grande adaptabilité et leurs bonnes performances sur l'analyse de séquences (Kim *et al.*, 2012b).

3. Données

Dans le cadre de notre travail, nous utilisons les données fournis par le *Centre pour L'édition Électronique Ouverte* (CLEO) extraites du Portail OpenEdition[13]. OpenEdition est une plateforme en ligne destinée à des articles électroniques, des livres, des blogs scientifiques dans le domaine des SHS. Il se compose de quatre sous plates-formes, Revues.org, Calenda, Books et Hypotheses. Le domaine des SHS est un domaine multidisciplinaire qui nous permet d'avoir une représentation de la grande majorité des types de formats des références bibliographiques. En effet, les conventions employées dans ce domaine sont très diversifiées par rapport à celles utilisées dans les sciences dites «dures». Bien qu'il existe des conventions plus largement utilisées comme celle de l'Association Américaine de Psychologie (*American Psychological Association*[14]) ou encore de l'Association du Langage Moderne (*Modern Language Association*[15]) de nombreuses organisations établissent leur propre convention afin de répondre à leurs besoins.
Les données que nous utilisons proviennent de l'index SoLr d'OpenEdition au format XML. Dans le cadre d'un programme de R&D du laboratoire OpenEdition trois corpus ont été annotés manuellement suivant le formalisme de la TEI[16]. Le premier corpus est exclusivement destiné à l'apprentissage de modèles pour les références bibliographiques présentes dans les zones bibliographiques. Le second corpus se concentre uniquement sur les notes de bas de pages fortement normées. Et enfin, le troisième corpus est constitué de références bibliographiques diffuses. Tous ces corpus sont établis uniquement sur les articles extraits de la plateforme Revues.org. Concernant la constitution du corpus des références bibliographiques diffuses, les revues sélectionnées ont été choisies aléatoirement dans la base de données d'OpenEdition sans tenir compte de l'appartenance disciplinaire de la revue ou du style bibliographique de la référence.
Dans le cadre de notre travail, nous nous intéressons uniquement au troisième cor-

13. `http://www.openedition.org/`
14. `http://www.apa.org/`
15. `https://www.library.cornell.edu/research/citation/mla`
16. TEI est un consortium qui développe, définit et maintient un langage de balisage pour décrire les caractéristiques structurelles et conceptuelles de textes.

pus orienté sur la détection des références bibliographiques diffuses[17]. Ce corpus se décompose en trois sous catégories, à savoir :

- **Notes de bas de pages particulières** : composée de 43 références bibliographiques, ces références sont des cas particuliers que l'on retouve dans les notes de bas de pages.

```
<note> Qui a cependant été, ensuite, jugée d'une sévérité excessive envers la bourgeoisie protestante
du xviie siècle. Ainsi, <bibl><author><forename>E.</forename> <surname>Labrousse</surname></author>
 montre qu'à partir du moment où la logique absolutiste était en marche, quelle que soit la stratégie
mise en œuvre par les protestants, elle se retournait contre eux (Cf. <title>« Une foi, une loi, un roi »
 ? La révocation de l'Édit de Nantes</title>, <pubPlace>Paris-Genève</pubPlace>,
<publisher>Payot-Labor</publisher> et <publisher>Fides</publisher>, <date>1985</date>)</bibl>.</note>
```

Figure 1 – Exemple de référence bibliographique de la catégorie Notes de bas de pages particulières

- **Références bibliographiques courtes** : composée de 449 références bibliographiques, cette catégorie réfère à des références composées de très peu d'éléments (de nombreuses variations entre auteur/date, titre/auteur, titre/page, etc.)

```
La pragmatique, qui a marqué le paysage des études de textes au cours des dernières décennies, se fonde
 sur un retour de l'intentionnalité, et conjointement, sur le rôle du contexte de la communication
(<bibl><author><surname>Strawson</surname></author> <date>1970</date></bibl>).
```

Figure 2 – Exemple de référence bibliographique de la catégorie Références bibliographiques courtes

- **Références bibliographiques entrelacées** : composée de 553 références bibliographiques, cette catégorie se compose de références qui sont «implicites», c'est-à-dire, qu'elles sont exprimées de manière informelle.

```
Dans <bibl>le <biblScope>premier numéro</biblScope> de l'<title>Almanach des gourmands</title>
(<author><forename>Grimod</forename> <surname>de la Reynière</surname></author>, <date>1804</date>
cité in <bibl> <author><surname>Duhart</surname></author> <date>2001</date></bibl>)</bibl>, il est
ainsi indiqué qu'il est possible de se procurer des pâtés de foie gras auprès des boutiques de
comestibles les plus réputées, telles que Corcellet, Rouget et l'Hôtel des Américains (Duhart, 2001).
```

Figure 3 – Exemple de référence bibliographique de la catégorie Références bibliographiques entrelacées

Suite à l'étude des références présentes au sein des revues constitutives du corpus des références bibliographiques diffuses nous avons identifié différents types de références ce qui explique les proportions différentes de références pour chacune des sous-catégories.

4. Méthodologie

Cette section présente, premièrement, le modèle de classification mis en place afin de détecter les zones contenant potentiellement des références bibliographiques. Et

17. `http://lab.hypotheses.org/`. Écrire à marin.dacos@openedition.org pour avoir accès à la collection.

secondement, les différents CRFs établis pour la détection des zones bibliographiques et l'annotation de leurs contenus.

4.1. *Classification supervisée des paragraphes contenant des références bibliographiques*

Au vu de la quantité importante de paragraphes ne comportant pas de références bibliographiques nous décidons d'effectuer un pré-filtrage via l'utilisation d'une classification supervisée. Pour ce faire, nous établissons deux classes : zone bibliographique/zone non bibliographique. La catégorie «Notes de bas de pages particulières» comporte 4,3% de paragraphes avec des zones bibliographiques sur un total de 306 paragraphes. La catégorie «Références bibliographiques courtes» contient 20,8% de paragraphes avec des zones bibliographiques sur un total de 725 paragraphes. Et la catégorie «Références bibliographiques entrelacées» dénombre 23,3% de paragraphes avec des zones bibliographiques sur un total de 1342 paragraphes. Ensuite, nous choisissons d'utiliser une technique de classification utilisant SVM. Pour l'implémentation du SVM, nous utilisons l'outil *SVMLight*[18] développé par (Thorsten, 1999). Concernant les paramétrages effectués, nous établissons une liste de mots les plus caractéristiques de chaque classe que nous utilisons comme attributs. Cette liste est réalisée grâce à l'algorithme *InfoGainAttribute* (IGA). IGA permet d'effectuer un ratio d'informations pour acquérir les informations intrinsèques, il est utilisé pour réduire un biais en faveur des attributs à valeurs multiples. Après plusieurs tests, nous avons choisi d'utiliser une fréquence minimale d'apparition des termes de 1 combinée à une liste pour laquelle nous avons supprimé les mots dont le score « Élimination récursive de caractéristiques »[19] (RFE) est égal à 0. Pour chaque sous-catégorie, nous effectuons 10 validations croisées afin d'évaluer la façon dont les résultats se généralisent à un ensemble de données. Le tableau 1 présente les résultats obtenus suite à la classification pour chacune des sous-catégories.

Nom corpus	accuracy	précision	rappel
Notes de bas de pages particulières	59,1%	66,7%	36,4%
Références bilbiographiques courtes	74,8%	71,0%	83,1%
Références bibliographiques entrelacées	79,0%	88,6%	71,2%

Tableau 1 – Résultats pour la classification supervisée

Concernant les performances de classification, pour la catégorie «Notes de bas de pages particulières», nous obtenons une «accuracy» sur le jeu de test de 59,1% ce

18. `http://svmlight.joachims.org/`
19. *Recursive Feature Elimination* : élimination récursive de caractéristiques.

qui correspond à 13 références classées correctement et 9 références classées comme incorrectes. Pour la catégorie «Références bibliographiques courtes», nous obtenons une «accuracy» sur le jeu de test de 74,8% ce qui correspond à 89 références classées correctement et 30 références classées comme incorrectes. Et pour la catégorie «Références bibliographiques entrelacées», nous obtenons une «accuracy» sur le jeu de test de 79,0% ce qui correspond à 156 références classées correctement et 44 références classées comme incorrectes. Nous pouvons noter que les résultats les plus faibles sont obtenus pour la catégorie «Notes de bas de pages particulières» ce qui s'explique par la faible quantité de paragraphes comportant des références bibliographiques (seulement 4,3%). Les deux autres catégories obtiennent des résultats similaires ce qui s'explique par une proportion égale de paragraphes comportant des zones bibliographiques comparativement à ceux n'en comptenant pas.

4.2. *Utilisation des CCRFs pour l'annotation des références bibliographiques*

Pour construire nos CCRFs, nous utilisons plusieurs caractéristiques afin de construire nos vecteurs d'informations. Les principales caractéristiques exploitées dans la littérature pour l'annotation automatique de références bibliographiques sont fondées sur un certain nombre d'observations tel que des caractéristiques lexicales ou morphologiques, à la fois, sur les champs et sur les mots contenus dans les champs. Afin d'élargir notre étude, nous avons également étudié les caractéristiques utilisées dans la détection des entités nommées qui est une tâche connexe. Faisant un parallèle entre la tâche de détection des entités nommées et l'analyse des références bibliographiques, nous sommes en mesure d'extraire des informations plus utiles dans la caractérisation des champs et des mots contenus dans les champs. La typologie des caractéristiques mise en place est la suivante :

1) **Caractéristiques contextuelles** : exploitation des mots ainsi que des catégories syntaxiques précédant et suivant le mot courant. Suite à des expériences simples, nous retenons comme caractéristiques, les trois mots précédents et suivants le mot courant ainsi que les deux catégories syntaxiques précédents et suivants le mot courant.

2) **Caractéristiques locales** : elles se divisent en quatre sous-catégories : les caractéristiques morphologiques, locationnelles, lexicales et syntaxiques. Les caractéristiques morphologiques permettent de définir la forme des mots (Ex : majuscule/minuscule). Les caractéristiques locationnelles définissent la position des champs dans la séquence. Les caractéristiques lexicales correspondent à l'exploitation de listes de mots prédéfinis. Enfin, les caractéristiques syntaxiques réfèrent à la ponctuation.

À partir de ces caractéristiques nous construisons des vecteurs d'informations pour chaque mot. La combinaison de ces caractéristiques nous permet d'établir des vecteurs d'informations robustes pour la définition de nos différents champs. La construction des CRFs est réalisée par l'outil *Wapiti*[20] développé par (Lavergne *et al.*, 2010). Nous

20. `http://wapiti.limsi.fr`

utilisons un jeu de 13 labels établi en fonction des différents champs rencontrés lors de l'études des références bibliographiques. Le tableau 2 présente les labels employés.

Type	Label	Description
Bibliographie	bibl	zone bibliographique
Auteur	surname forename	nom de famille prénom
Titre	title	titre de l'article visé
Éditeur	publisher	éditeur
Date	date	date, la plupart des années
Place	place	nom de lieux
Édition	biblscope	informations sur les pages, volume, numéro, etc.
Détail	extent edition	nombre total de pages autres renseignements descriptifs de l'édition
Etc.	abbr punc ref	abbréviation ponctuation url

Tableau 2 – Descriptions des labels

Suite à la classification, nous obtenons une liste des paragraphes contenant potentiellement des références bibliographiques. Le premier CRF permet d'effectuer un premier filtrage afin d'identifier la zone dans laquelle les différents champs bibliographiques sont situés. Ce premier CRF permet de réduire le champ d'exécution du deuxième CRF via l'apposition de la balise <bibl> autour de la séquence contenant des champs bibliographiques. Le deuxième CRF nous permet d'annoter le contenu de la séquence détectée par le premier CRF via l'identification des différents champs bibliographiques, tels que, l'auteur, le titre, etc.

5. Expérimentations et résultats

Dans la section suivante, nous présentons les différentes expérimentations effectuées ainsi que les performances obtenues lors de la détection des zones bibliographiques et l'annotation de leurs contenus.

5.1. *Détections des zones bibliographiques*

Cette section présente les résultats obtenus suite à l'apposition de la balise <bibl> autour des séquences contenant des champs bibliographiques. Pour effectuer nos ex-

périmentations, nous établissons un corpus composé uniquement de paragraphes comprenant des références bibliographiques (Corpus de référence), c'est à dire, que ce corpus est établi indépendamment de la classification présentée précedemment. Ce qui nous donne pour la catégorie «Notes de bas de pages particulières» 30 paragraphes contenant des références bibliographiques, pour la catégorie «Références bibliographiques courtes» 158 paragraphes contenant des références bibliographiques et pour la catégorie «Références bibliographiques entrelacées» 311 paragraphes contenant des références bibliographiques. Nous avons fait ce choix afin de mesurer la robustesse des caractéristiques que nous avons choisies afin de constituer nos vecteurs d'informations. Nous présentons des expérimentations basées sur 10 validations croisées composées de 70% de corpus d'apprentissage et de 30% de corpus de test afin d'éviter un sur-apprentissage dû à une quantité trop importante de données d'apprentissage.

Nom corpus	Corpus de référence		
	précision	rappel	F-mesure
Notes de bas de pages particulières	72,2%	73,4%	72,8%
Références bilbiographiques courtes	94,1%	93,4%	**93,8%**
Références bibliographiques entrelacées	70,2%	72,6%	71,4%

Tableau 3 – Résultats pour la détection des zones bibliographiques

Le tableau 3 présente les résultats obtenus suite à la détection des zones bibliographiques pour chacune des sous-catégories. Nous observons que les résultats les plus faibles sont obtenus par la catégorie «Références bibliographiques entrelacées» et les résultats les plus significatifs par la catégorie de «Références bilbiographiques courtes». Ce phénomène peut être expliqué par la complexité de la structure de ces dernières mais également par un degré d'implicite plus important que les références courtes. En effet, les «Références bilbiographiques courtes» ont une structure beaucoup plus formelle que les «Références bibliographiques entrelacées». Les références courtes sont généralement utilisées afin de citer, de rapporter les mots ou les phrases de quelqu'un. Ce type de références est structuré plus formellement et répond à des conventions plus strictes tandis que les références entrelacées se fondent dans le corps du texte. Nous notons également que les catégories «Références bibliographiques entrelacées» et «Notes de bas de pages particulières» obtiennent substanciellement les mêmes F-mesures avec cependant une précision légéremment supérieure pour la catégorie «Notes de bas de pages particulières». Ce phénomène s'explique par la structuration similaire des références présentes au sein de ces deux catégories ainsi qu'un degré similaire d'implicite. Cette analyse nous permet de rendre compte de la difficulté qui réside dans la détection des zones bibliographiques plus le degré d'implicite est important. En effet, le degré d'implicite des références bibliographiques engendre des structurations et des répartitions dans le texte très variées ce qui rend l'homogénéisation des traitements difficile.

5.2. *Détection des champs bibliographiques*

Dans cette section, nous présentons les résultats obtenus suite à l'identification des différents champs bibliographiques. Pour cette expérimentation nous choisissons de présenter les résultats obtenus d'une part, sur un apprentissage effectué sur des modèles directement établi en fonction de chacune des sous-catégories et d'autre part, sur un apprentissage effectué sur un modèle établi sur le premier corpus[21] réalisé par le laboratoire OpenEdition. Pour rappel, le premier corpus d'OpenEdition est exclusivement destiné à l'apprentissage de modèles pour les références bibliographiques présentes dans les zones bibliographiques. Nous choisissons d'établir cette comparaison afin de constater l'impact des performances d'un modèle appris sur des données fortement structurées et de son application sur des données de nature différente. Cette analyse nous permet également de comparer notre approche aux approches qu'emploient les outils basés sur des CRFs présentées précédemment dans la section 2. Nous précisons également que cette expérimentation est faite indépendamment des résultats obtenus lors de la détection des zones bibliographiques. Les corpus utilisés ne découlent donc pas directement des résultats présentés lors de l'expérimentation précédente. Nous présentons des expérimentations basées sur 10 validations croisées composées de 70% de corpus d'apprentissage et de 30% de corpus de test afin d'éviter un sur-apprentissage dû à une quantité trop importante de données d'apprentissage.

Nom corpus	Corpus de référence		
	précision	rappel	F-mesure
Notes de bas de pages particulières	86,2%	80,1%	82,9%
Références bilbiographiques courtes	89,1%	88,8%	**89,1%**
Références bibliographiques entrelacées	79,7%	65,6%	71,8%
Notes de bas de pages particulières Corpus 1	69,1%	62,4%	65,5%
Références bilbiographiques courtes Corpus 1	69,0%	73,8%	71,2%
Références bibliographiques entrelacées Corpus 1	34,2%	27,5%	30,5%

Tableau 4 – Résultats pour la détection des champs bibliographiques

Le tableau 4 nous permet d'observer une nette dégradation des performances suite à l'apprentissage effectué sur le premier corpus d'OpenEdition. Ceci s'explique par la structuration très formelle que l'on retrouve dans les références bibliographiques présentes dans les zones bibliographiques. En effet, ce type de références répondent à des conventions très strictes. Il est intéressant de noter que les références courtes

21. Dans le tableau 4 se corpus est présenté avec l'extension Corpus 1

obtiennent les meilleures performances suite à l'apprentissage sur le premier corpus d'OpenEdition. Nous expliquons ce phénomène par le fait que ce type de références est structuré plus formellement et répond à des conventions plus strictes à la différence des deux autres catégories. Cette expérience nous permet de mettre en évidence une perte importante de performance lorsque l'on change le contexte applicatif sur lesquels les modèles ont été appris. Concernant les performances obtenues suite à l'apprentissage effectué sur des modèles directement établi en fonction de chacune des sous-catégories, les résultats les plus faibles sont obtenus pour la catégorie «Références bibliographiques entrelacées». Ce phénomène peut s'expliquer par la diversité des champs bibliographiques que l'on peut trouver dans cette catégorie associé à un degré d'implicite important. À l'inverse, la catégorie «Références bilbiographiques courtes» obtient les meilleurs résultats ce qui est due à la présence de champs bibliographiques plus récurrent. En effet, nous dénombrons une moyenne de cinq champs bibliographiques différents pour la catégorie «Références bilbiographiques courtes» tandis que pour la catégorie «Références bibliographiques entrelacées» nous recensons douze champs bibliographiques différents.

6. Conclusion

En conclusion, nous avons pu observer des résultats très variables concernant la classification supervisée. Ce phénomène est induit par les quantités de données d'apprentissage très inégales selon les sous-catégories suite à l'échantillonnage aléatoire des revues pour établir les différents corpus. Concernant la détection des zones bibliographiques, nous avons pu rendre compte de la difficulté qui réside dans la détection des zones bibliographiques plus le degré d'implicite est important. Nous avons pu noter que le degré d'implicite des références bibliographiques engendre des structurations et des répartitions dans le texte très variées ce qui rend l'homogénéisation des traitements difficile. Concernant la détection des champs bibliographiques, les expériences effectuées ont permis de mettre en evidence une perte importante de performance lorsque l'on change le contexte applicatif sur lesquels les modèles ont été appris. Lors de cette même étude, nous avons pu observer que les champs bibliographiques qui composent les références bibliographiques sont très variés selon les sous-catégories. Les différentes expériences menées sur les références bibliographiques diffuses nous permettent de constater d'une part, que l'identification des zones bibliographiques est une tâche complexe selon le degré d'implicite de la référence et d'autre part que, la nature des champs bibliographiques qui composent ces références bibliographiques diffuses est très variés.
Dans nos futurs travaux, nous comptons trouver une autre granularité que les paragraphes afin de réduire le champ d'action des CRFs. Nous envisageons d'établir des corpus mieux proportionnés mais également plus représentatifs de tous les cas possibles. Nous comptons choisir des caractéristiques spécifiques à chaque sous-catégorie pour améliorer les performances lors de l'annotation que ce soit pour améliorer la détection des zones bibliographiques que pour l'annotation de leurs contenus. À court terme, notre but s'oriente sur la réalisation d'une chaîne de traitement complète dans

lequelle nous allons intégrer un SVM multiclasse qui guidera directement le choix du premièr CRF. Nous rappelons que la finalité de ce travail consiste en la mise en œuvre d'un système de recommandation de lectures dans lequel nous comptons utiliser les références bibliographiques afin d'établir des liens entre les documents.

7. Bibliographie

Anzaroot S., McCallum A., « A new dataset for fine-grained citation field extraction », *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

Bellis N. D., *Bibliometrics and citation analysis : from the science citation index to cybermetrics*, Scarecrow Press, 2009.

Chen C.-C., Yang K.-H., Chen C.-L., Ho J.-M., « Bibpro : A citation parser based on sequence alignment », *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, n° 2, p. 236-250, 2012.

Councill I. G., Giles C. L., Kan M.-Y., « ParsCit : an Open-source CRF Reference String Parsing Package. », LREC, p. 28-30, 2008.

Cronin B., *The Citation Process : The Role and Significance of Citations in Scientific Communication*, Taylor Graham, London, 1984.

Deconinck S., « Artificial intelligence a modern approach », 2010.

Giles C. L., Bollacker K. D., Lawrence S., « CiteSeer : An automatic citation indexing system », *Proceedings of the third ACM conference on Digital libraries*, ACM, p. 89-98, 1998.

Huang I.-A., Ho J.-M., Kao H.-Y., Lin W.-C., « Extracting citation metadata from online publication lists using BLAST », *Advances in Knowledge Discovery and Data Mining*, Springer, p. 539-548, 2004.

Kim Y.-M., Bellot P., Faath E., Dacos M., « Automatic annotation of bibliographical references in digital humanities books, articles and blogs », *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, ACM, p. 41-48, 2011.

Kim Y.-M., Bellot P., Faath E., Dacos M., « Annotated Bibliographical Reference Corpora in Digital Humanities. », LREC, p. 329-340, 2012a.

Kim Y.-M., Bellot P., Tavernier J., Faath E., Dacos M., « Evaluation of BILBO reference parsing in digital humanities via a comparison of different tools », *Proceedings of the 2012 ACM symposium on Document engineering*, ACM, p. 209-212, 2012b.

Lavergne T., Cappé O., Yvon F., « Practical very large scale CRFs », *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 504-513, 2010.

Lopez P., « GROBID : Combining automatic bibliographic data recognition and term extraction for scholarship publications », *Research and Advanced Technology for Digital Libraries*, Springer, p. 473-474, 2009.

Rezaei B. A., Muntz A. H.-Y. M., « System and method for context-based knowledge search, tagging, collaboration, management, and advertisement », February 19, 2013. US Patent 8,380,721.

Thorsten J., Making large scale SVM learning practical, Technical report, Universität Dortmund, 1999.