Extraction de citations contenues dans des documents brevet

Eleni Kogkitsidou, Tita Kyriacopoulou, Claude Martineau, Cristian Martinez, A-Young Kim, Antoine Schoen

Extraction de citations contenues dans des documents brevet

Eleni Kokgitsidou[1], Tita Kyriacopoulou[2], Claude Martineau[2], Cristian Martinez[1], Kim A-Young[1], Antoine Schoen[1],
[1] École Supérieure d'Ingénieurs en Électrotechnique et Électronique/IFRIS
[2] Université Paris-Est Marne-La-Vallée/LIGM

Résumé : Le présent article s'inscrit dans une démarche générale d'élaboration d'outils et de méthodes d'analyse permettant de caractériser les activités scientifiques et techniques. Le nombre de publications scientifiques numériques est de plus en plus important. Nous nous intéressons plus particulièrement ici au repérage et à l'extraction automatique de citations et de références contenues dans des documents, en anglais, de type brevet d'inventions. La méthode utilisée repose sur une approche symbolique qui fait appel à la création et l'utilisation combinée de dictionnaires électroniques et de grammaires locales. L'outil de traitement de corpus Unitex est utilisé pour l'élaboration et l'application de ces ressources linguistiques à un corpus d'étude.

Keywords: extraction d'information, citations, brevets, grammaires locales, Unitex.

1. Introduction

La publication scientifique en général étant de plus en plus importante sous une forme numérique, un traitement automatique de ce type de textes s'avère de plus en plus nécessaire. Le travail présenté ici porte sur l'extraction automatique de citations et de références dans des documents de type brevet. Elle est une étape dans l'analyse des relations entre les publications scientifiques et le dépôt de brevets. Notre approche repose sur une méthode symbolique qui fait appel à la création et l'utilisation combinée de dictionnaires électroniques et de grammaires locales.

La présente étude s'inscrit dans le cadre d'une synergie entre les domaines de l'étude des sciences et technologies et de la linguistique informatique. Elle est le fruit d'une collaboration entre l'IFRIS[1] de l'ESSIE[2] et le LIGM[3] de L'Université Paris-Est Marne-la-Vallée. L'équipe Pôle Indicateurs de l'IFRIS, créée en 2009 a pour objectif la collecte et le traitement des informations scientifiques et techniques afin de nourrir des analyses sur les dynamiques de production de la connaissance. Notre exposé se décompose en cinq parties. Nous présentons tout d'abord notre corpus de travail et le type d'extractions à effectuer (section 2), puis les méthodes utilisées pour élaborer les annotations manuelle et automatique (section 3). Nous exposons ensuite (section 4) les problèmes rencontrés lors du traitement du corpus étudié puis nous procédons (section 5) à une évaluation des extractions fournies. Nous évoquons enfin les perspectives que nous ouvre ce travail.

2. Problématique et corpus d'étude

Notre corpus de travail, est composé par des documents brevet, en anglais, concernant le domaine de l'oncothérapie qui peuvent être téléchargés gratuitement sur le site Espacenet[4]. La sélection de brevets à été réalisée par l'équipe du Pôle Indicateurs. Le corpus comporte presque 1 600 000 mots en 140 000 lignes, et environ 2 530 pages. Le corpus est divisé en 77 fichiers, un par brevet. Nous essaierons d'exploiter les formalismes et les types de citations contenues dans les documents brevet ainsi que la méthodologie à suivre pour effectuer l'annotation d'un corpus de brevets.

Les documents brevet que nous devons traiter se présentent sous la forme de texte en format libre (.txt) et en aucun cas des formats préétablis, de type formulaire par exemple. Ainsi le terme "citation" est utilisé pour désigner les chaînes de caractères qui sont disposées librement dans un texte, dans un ordre, un format, et une ponctuation variés. Les citations font le lien entre le texte et un élément bibliographique particulier. A ce titre, elles permettent de l'identifier et de le localiser. Une référence est à son tour composée d'un ensemble de métadonnées défini, telles que, par exemple, les noms d'auteurs, le titre du document, le lieu de publication, la date de publication, etc. Une citation établit donc un lien direct entre une référence bibliographique et plusieurs citations, ces dernières pouvant renvoyer à une référence unique.

2.1. Type de citations extraites

Les citations dans les documents brevet se divisent en deux types : les *patent literature* (ou *brevet*) et les *non-patent literature* (ou *non brevet*). Voici tout d'abord un exemple de citations du type *non brevet* :

- Nishisho et al., 1991
- Kakiuchi et al, (2004) Hum Mol Genet.; 13:3029-43 & (2003) MoI Cancer Res.; 1:485-99

[1] Institut Francilien Recherche Innovation et Société : http://ifris.org
[2] École Supérieure d'Ingénieurs en Électrotechnique et Électronique
[3] Laboratoire d'Informatique Gaspard-Monge : http://igm.univ-mlv.fr/LIGM/
[4] http://www.epo.org/searching/free/espacenet.html

Chaque citation est liée à une référence. Souvent, nous rencontrons les références des citations à la fin de chaque texte. Les références sont les "explications" des citations autrement dit elles présentent les citations d'une façon plus détaillée, voici un exemple :

Citation	Nishisho et al., 1991
Référence de la citation	Nishisho I, Nakamura Y, Miyoshi Y, Miki Y, Ando H, Horii A, KoyamaK, Utsunomiya J, Baba S and Hedge P. (1991). Science, 253,665-669.

Tableau 1: Exemple de citation

Les citations du type non brevet, se réfèrent essentiellement à des articles, mais elles peuvent également concerner d'autres entités extérieures telles que les bases de données, elles ont la tendance à être plus longues, moins strictement structurées que les citations brevet, et avec plus d'emphase sur les mots des contextes (Galibert 2010). Ce type de citations comporte des informations sur les auteurs, la date, les pages et le titre du livre ou du journal comme le montrent l'exemple du tableau 1.

Les citations de type *brevet* se présentent de la façon suivante :

- U. S. Pat. No. 5,223, 409
- U.S. Patent Nos. 6,150,584; 5,545,807; 5,545,806; 5,569,825; 5,625,126; 5,633,425; 5,661,016

Elles comportent souvent un code de pays ou organisation (*authority*) comme DE, FR, US ou un code de région comme EP ou WO, un numéro qui peut contenir une année, une date, des commentaires et un code de type (*kind code*).

2.3 Représentation des annotations

Les attributs possibles dans les citations non brevet sont le nom d'auteur, la date de publication, le titre du journal ou de la publication, le volume, les numéros de pages, le nom de l'éditeur, le nom de la maison d'édition, le lieu, et le titre de l'article. Nous avons défini un format d'annotation détaillée, utilisant ces attributs sous la forme d'un balisage de type XML.

Voici un exemple d'une citation non brevet annotée :

*<citation>***Loeffler and Behr, Methods Enzymol 217: 599-618, 1993***</citation*
author=Loeffler||lastname=Loeffler|author=Behr||lastname=Behr|journal=Methods Enzymol|
volume=217:|pages=599-618|date=1993>

Les attributs possibles dans les citations brevet sont l'autorité, le numéro, la date, les commentaires et le code de type.

Voici un exemple d'une citation brevet annotée :

*<patentref>***US Pat. Nos.5,091,513, 5,132,405, and 4,946,778***</patentref*
authority=US|number=5,091,513,5,132,405, 4,946,778>

3. Traitement du corpus

Pour analyser notre corpus nous avons utilisé Unitex[5], une plate-forme de traitement de corpus fondée sur des graphes à états finis et des RTN[6]. Développé principalement par S. Paumier (Paumier), cet outil permet de construire et gérer des ressources linguistiques telles que des dictionnaires et des grammaires, et de les appliquer à des textes.

3.1 Annotation manuelle

L'annotation manuelle a été réalisée de manière semi-automatique. Un premier traitement a consisté à appliquer aux documents du corpus une grammaire (cf. Figure 1) donnant une annotation approximative des citations et références. Cette grammaire n'utilise pas de dictionnaire. Elle utilise le « squelette » du motif recherché : séparateurs, initiales de prénoms suivis de points, volume, pages. L'annotation produite est ensuite corrigée et complétée manuellement.

[5] http://www-igm.univ-mlv.fr/~unitex/
[6] Recursive Transition Network

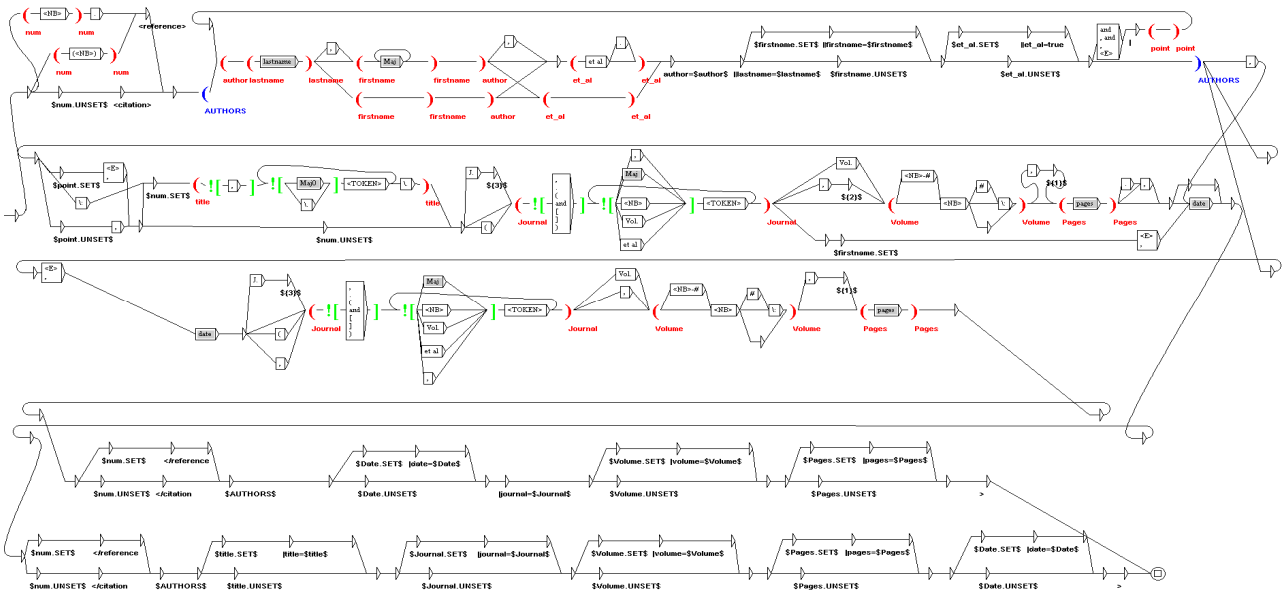

Figure 1: Graphe de pré-annotation manuelle

Les deux premiers tiers du graphe de pré-annotation manuelle (Figure 1) effectuent l'extraction des différentes parties constitutives potentielles d'une citation et les mémorisent dans des variables. Le tiers restant construit les couples attributs valeurs pour les parties présentes (variables non-vides) et les place dans une balise fermante </citation> avec le format: </citation attribut=valeur… attribut=valeur>.

3.2 Extraction et annotation automatique

L'annotation automatique a fait appel à la création de ressources de type dictionnaires et à une grammaire décrivant beaucoup plus précisément les différents motifs éventuels d'une citation ou référence. Notre grammaire est constituée de deux graphes principaux de complexité inégale. Le premier graphe *NPL* concerne le traitement des citations non brevet, le second *Patent* celui des citations brevet. Le graphe principal *citation* de la figure 2 fait appel à ces deux graphes et réalise l'annotation en construisant les balises.

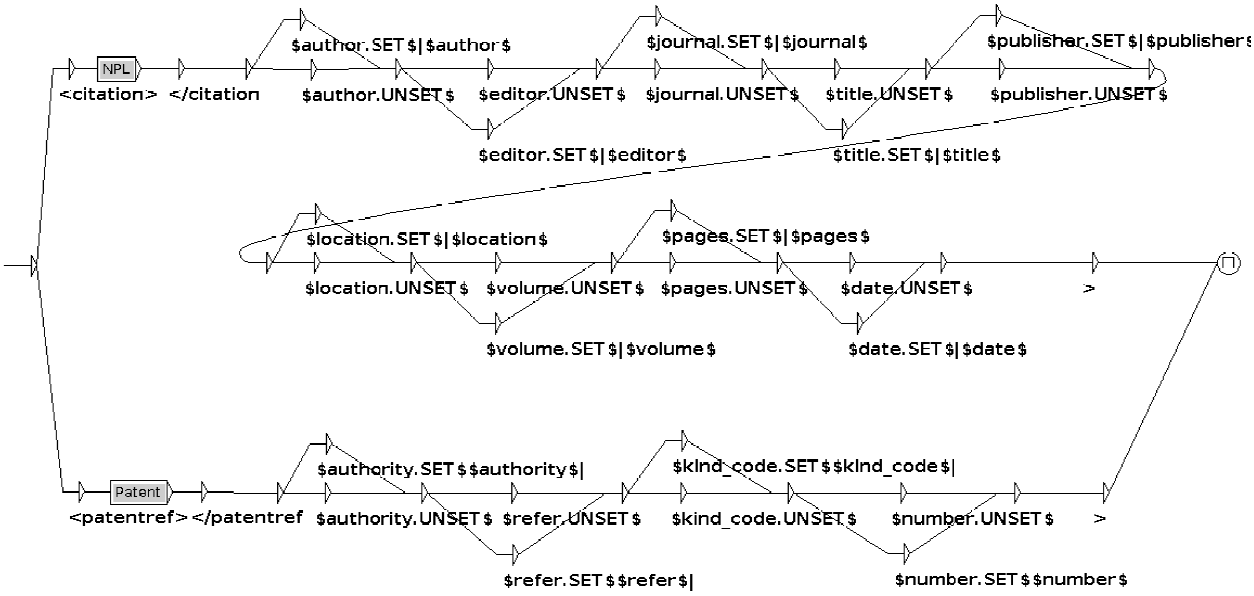

Figure 2: Graphe principal citation

Le sous-graphe *Patent* que voici :

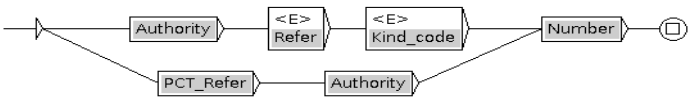

Figure 3: Graphe de citations brevet

reconnaît les citations de type brevet. Le sous-graphe *Authority* contient la liste des codes pays/organisation[7] officiels, et la liste des noms de pays. Les sous-graphes *Refer* et *PCT_Refer* reconnaissent les chaînes de caractères comme « patent », « patent application[8] », ou « PCT[9] application ». Le sous-graphe *Kind_code*[10] contient les deux listes des « Kind_code. Refer[9] » et « Kind_code » mais sa présence n'est pas obligatoire pour le repérage des citations brevet ('<E>' représente un token 'vide'). Le dernier sous-graphe *Number* force à reconnaître une chaîne de chiffres comprenant éventuellement des ponctuations. La figure 4 montre le résultat de reconnaissance de citations brevet pour les fichiers numéro 00-17 du corpus oncothérapie.

Figure 4: Exemples de citations brevet reconnues

Le graphe de la figure 5 reconnaît les citations non-brevet.

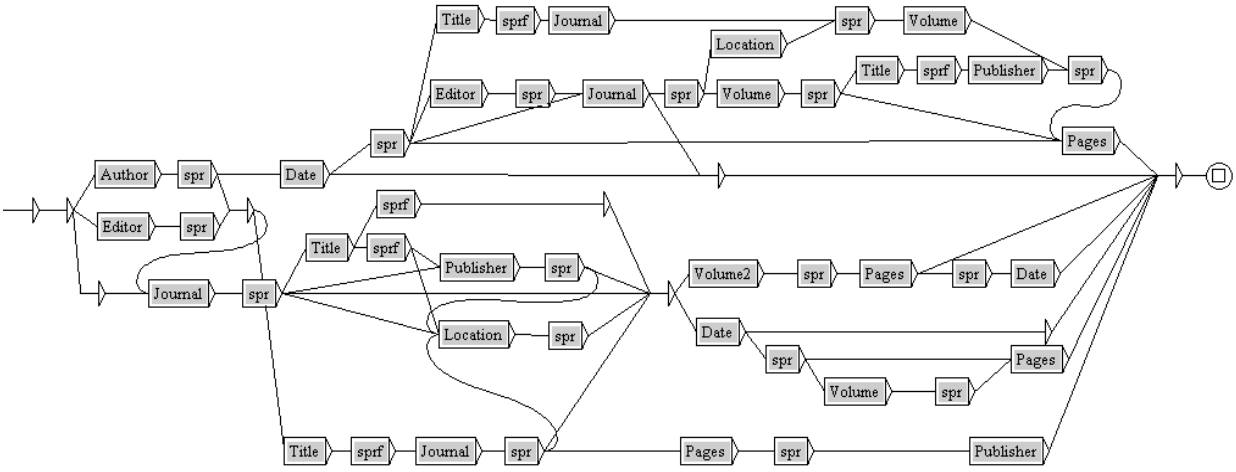

Figure 5: Graphe de citations non-brevet

La structure de ce graphe est beaucoup plus complexe que le graphe brevet, car les motifs des citations non brevet sont très variés. Ce graphe contient les motifs qui ont été fréquemment observés dans le corpus oncothérapie et le corpus CORA[11] qui sert de référence dans le domaine de la bibliographie.

Dans les sous-graphes *Author* et *Editor*, nous avons utilisé *DiTex-PersonName*, une ressource linguistique antérieurement développée dans le projet DiTex.org[12] pour la détection des noms de personne. DiTex-PersonName contient environ 256,000 prénoms uniques avec de traits de genre et 556,000 noms de famille uniques avec plusiers traits permettant de récupérer les composants du nom de famille : préfixes et suffixes et indiquer s'ils sont polysémiques.

[7] Les codes de pays sont composés de deux lettres (par exemple GB) indiquant le pays dans lequel, ou l'organisation auprès de laquelle, la demande de brevet a été déposée ou a donné lieu à la délivrance d'un brevet.
[8] Une demande de brevet
[9] Le Traité de coopération en matière de brevets ou PCT (pour *Patent Cooperation Treaty*), signé à Washington le 19 juin 1970 permet de simplifier les demandes de brevets et est reconnu dans 117 pays.
[10] Ces codes permettent de caractériser l'étape dans laquelle se trouve la demande de dépôt.
[11] http://people.cs.umass.edu/~mccallum/data.html
[12] http://ditex.org/

Ces ressources ont utilisées en tant que dictionnaires dans Unitex.La figure 6 montre le résultat d'une reconnaissance des citations non-brevet dans le corpus *oncothérapie*.

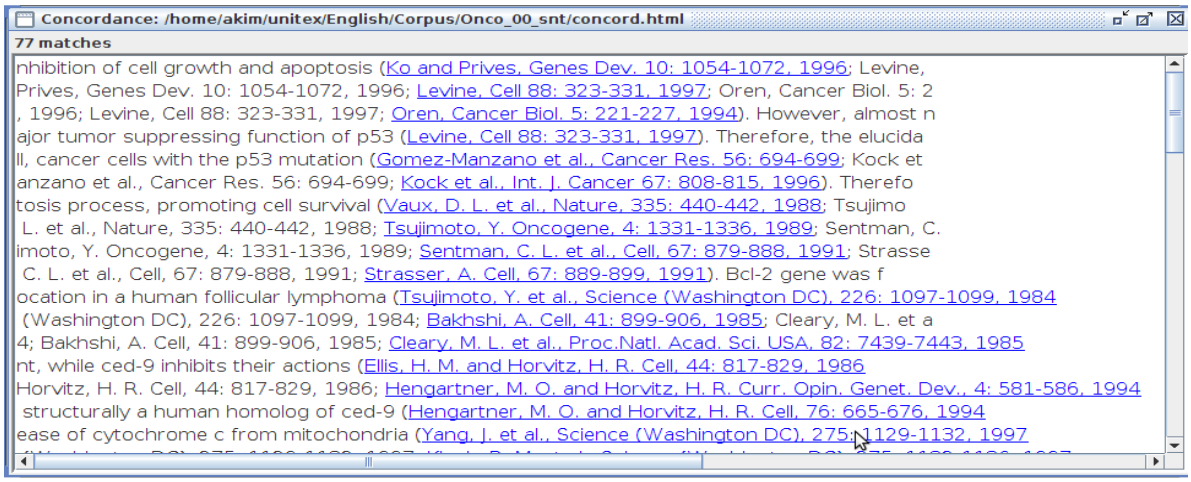

Figure 6: Exemples de citations non-brevet reconnues

Après l'application du graphe principal, nous avons rencontré des problèmes d'ambiguïtés. Nous avons observé par exemple les annotations ambiguës concernant le titre de la revue :

<citation>Levine, Cell 88: 323-331, 1997</citation |author=Levine|journal=Cell|volume=88|pages=323-331|date=1997>
<citation>Levine, Cell 88: 323-331, 1997</citation |journal=Levine|title=Cell|volume=88|pages=323-331|date=1997>

Tableau 2: Exemple de résultats ambigus

Cette difficulté nous a conduits à construire des ressources linguistiques. Ainsi le graphe *citations non-brevet* utilise entre autre, deux dictionnaires additionnels : le premier concerne les titres de journaux et le second les abréviations de titres de journaux. Pour construire ces dictionnaires, nous avons collecté les listes de titres des journaux et leurs abréviations depuis les ressources suivantes :

- CAplus Core Journal Coverage List (http://www.cas.org/content/references/corejournals)
- Serials Source List for Biological Sciences (http://www.csa.com/ids70/serials_source_list.php?db=biolclust-set-c)
- Serials Source List for METADEX (http://www.csa.com/ids70/serials_source_list.php?db=metadex-set-c)
- Science & Engineering Journal Abbreviations(http://scieng.library.ubc.ca/coden/)
- Aqualignt Journal Titles Abbreviations List (http://www.aqualight.info/journal_abbrevs/abbreva.htm)

Les graphes *citations brevet* et *citations non-brevet* ne produisent pas de sorties, mais conservent les valeurs de chaque élément reconnu dans les variables. Seul le graphe principal citation (Figure 2) qui fait appel aux deux graphes précédemment cités (Figure 3 graphe citations brevet et Figure 5 graphe de citations non brevet) produit des sorties sous forme de balise et construit ainsi l'annotation. La figure 7 montre un exemple des extractions produites.

<citation>Watanabe H, et al., (1998) Proc Natl Acad Sci</citation |author=Watanabe H||et_al=true|journal=Proc Natl Acad Sci|date=1998>
<citation>Williams et al., (1997)</citation |author=Williams||et_al=true|date=1997>
<citation>Wrighton et al, (1996) Science 273: 458-64</citation |author=Wrighton||et_al=true|journal=Science|volume=273|pages=458-64|date=1996>
<citation>Yano Y, et al, (2004) Cancer Lett.;207: 139-47</citation |author=Yano Y||et_al=true|journal=Cancer Lett.|volume=207|pages=139-47|date=2004>
<citation>Yao M, et al, (2005) J Pathol.;205:377-87</citation |author=Yao M||et_al=true|journal=J Pathol.|volume=205|pages=377-87|date=2005>
<citation>Zuckermann et al. (1994) J. Med. Chem. 37: 2678-85</citation |author=Zuckermann||et_al=true|journal=J. Med. Chem.|volume=37|pages=2678-85|date=1994>
<citation>van der Bruggen, (1996) J Exp Med 183: 725-9</citation |author=van der Bruggen|journal=J Exp Med|volume=183|pages=725-9|date=1996>
<patentref>US Pat. Application 2002103360</patentref authority=US|refer=Pat. Application|number=2002103360>
<patentref>US Pat. No. 5,223,409</patentref authority=US|refer=Pat.|number=No. 5,223,409>
<patentref>US Pat. No. 5,571,698; 5,403,484, and 5,223,409</patentref authority=US|refer=Pat.|number=No. 5,571,698|5,403,484|5,223,409>
<patentref>US Patent No. 6,506,559</patentref authority=US|refer=Patent|number=No. 6,506,559>
<patentref>WO2004076623</patentref authority=WO|number=2004076623>

Figure 7: Exemples d'annotations produites.

4. Problèmes rencontrés

Pendant la procédure de l'annotation manuelle nous avons rencontré divers problèmes, qui sont explicités ci-dessous.

4.1. Ambiguïtés entre les catégories

Un des problèmes fréquent était celui de la casse : nom du journal écrit en majuscules et qui se confond, de ce fait, avec le nom de l'auteur.

FASEB J (1992) 6,2422-2427)

Pour résoudre ce problème il était indispensable de faire une recherche approfondie sur le web afin de définir s'il s'agit d'un nom d'auteur ou d'un nom de journal.

Une autre source d'ambigüité était liée au fait de déterminer les frontières des sous-motifs présents dans une citation en fonction de la présence ou non de séparateurs clairs entre ces sous-motifs. Par exemple « J. » est-il le second prénom d'un auteur (appartenant au sous-motif auteur) ou l'abréviation de la mention « Journal » appartenant au sous-motif de même nom. De même, comment trouver la fin du titre qui est un motif libre. Dans ce dernier cas la recherche et l'attente d'un point ou d'un guillemet semblent une bonne approche.

4.2. Citations multiples

Dans plusieurs cas, deux citations se réfèrent au même auteur et sont coordonnées par le symbole "&" ou par le littéral « and ».

Kakiuchi et ah, (2004) Hum MoI Genet.; 13:3029-43 **&** (2003) MoI Cancer Res.; 1:485-99

Nous avons convenu d'une étiquette particulière pour annoter ce type de citation.

4.3. Mauvaises transcriptions et synonymes graphiques

Sur certains fichiers du corpus nous avons observé plusieurs mauvaises transcriptions comme la substitution de la lettre "a" par le mot "alpha" et la substitution des synonymes graphiques "5" par "S et "h" par "l" :

Mark DF et **[alpha]**l., (1984) Proc Natl Acad Sci USA 81: 5662-6.;

Ishida I**5** et a**h**, Cloning and Stem Cells 4: 91-102, 2002

Avec des substitutions systématiques nous avons réussi à homogénéiser notre corpus.

4.4. Déclaration de pages au milieu de citations

Les premiers résultats (des motifs extraits par les applications des grammaires locales) obtenus comportaient des déclarations de pages :

Kawano et al, Cancer Res 60: 3550-8 **[0012]** (2000)

Kubo et al. , J Immunol 152: **<Desc/Clms Page number 4>** 3913-24 (1994)

Quand la déclaration de page coïncidait au milieu d'une citation, l'annotation repérée était endommagée. Pour cette raison nous avons éliminé toutes les déclarations de pages.

5. Evaluation de l'annotation automatique

Les systèmes de recherche d'information sont traditionnellement évalués en termes de rappel et précision (Grouin, 2011). Nous n'avons pas dérogé à cette règle et nous avons donc utilisé ces métriques ainsi que la F-mesure.

Rappelons les définitions :

$$Rappel = \frac{\text{Nombre de citations/références } \mathbf{correctement\ reconnues}}{\text{Nombre de citations/références } \mathbf{existantes}}$$

$$Précision = \frac{\text{Nombre de citations/références } \mathbf{correctement\ reconnues}}{\text{Nombre de citations/références } \mathbf{reconnues}}$$

La F-mesure (ou F-score) est une pondération qui combine la précision et le rappel :

$$F_\beta = \frac{(1+\beta) \text{ x } Précision \ x \ Rappel}{\beta^2 \text{ x } Précision + Rappel} \quad (\beta > 0)$$

En ce qui concerne le repérage des citations en premier lieu nous avons besoin d'effectuer deux détections :

- une au niveau de *l'étiquette* à travers des bornes fixées automatiquement (frontières de l'ensemble de la citation) ;
- et en second lieu, une autre au niveau de la classe associée à chaque *attribut* (type d'attribut).

Pour exécuter l'évaluation nous avons créés trois scripts perl. Afin de représenter les résultats de comparaison entre les appariements. Nous avons utilisé les étiquettes suivantes :

corrects	("==")
manquants	("+-")
faux positifs	("-+")
partiellement corrects	("~~")
incompatibles	("<>")

Tableau 3 : Résultat obtenus

De cette façon nous avons obtenu dans un fichier de résultat les appariements référence et hypothèse sous la forme :

« annotation du corpus de référence », « étiquette », « annotation automatique »

Grâce aux appariements nous avons pu faire les calculs de Precision, Rappel et F-score au niveau de l'étiquette et de l'attribut. Comme le tableau 2 montre les résultats de ces mesures pour les fichiers du corpus annoté, nous constatons que les colonnes qui figurent comportent trois valeurs : *Stricte* pour les appariements corrects ("=="), *Partiel* pour l'ensemble des appariements partiellement corrects ("~~") et corrects ("==") et *Moyen* pour l'ensemble de ces deux calculs.

	Etiquette			**Attribut**			**Etiquette & Attribut**
	Stricte	**Partiel**	**Moyen**	**Stricte**	**Partiel**	**Moyen**	**Moyenne Generale**
Précision	81.72%	88.08%	84.9%	80.29%	86.33%	83.31%	84.11%
Rappel	76.97%	84.33%	80.63%	61.62%	71.12%	66.37%	73.51%
F-score	79.45%	85.19%	82.7%	70.99%	77.42%	74.84%	78.43%

Tableau 4: Évaluation des résultats

Nous observons que les résultats de la précision (84.11%) sont bons en comparaison d'autres études basées sur l'extraction de citations par les brevets comme celui de Galibert (2010) qui présente le résultat général pour tags et tags spécifiques pour la précision (55.3%). Pourtant le rappel au niveau d'attributs semble insuffisant. Ce problème est dû à la présence des références bibliographiques qui ne sont pas traitées par les grammaires locales, la présence de mauvaises transcriptions dans certains fichiers annotés automatiquement et la substitution de synonymes graphiques "5" par "S et "h" par "l". De plus l'annotation automatique a été perturbée par la déclaration du numéro de pages au milieu de citations (*cf.* section 4).

Les résultats pourront s'améliorer avec le perfectionnement des grammaires locales et la normalisation du corpus (suppression de numéros de pages et substitution automatique de mauvaises transcriptions et de synonymes graphiques).

Conclusion

Cette étude avait comme but la présentation d'une méthode de repérage annotation d'un corpus de citations dans des documents de type brevet et la définition d'un protocole d'évaluation pour mesurer la qualité de l'annotation produite. Les résultats obtenus s'avèrent encourageants et nous permettent d'envisager une extension de la méthode appliquée à une plus grande variété de documents, les formats de citations et de références n'étant ni stables ni normalisés. En outre, les évolutions récentes de l'environnent Unitex nous donnent la possibilité d'effectuer un passage à l'échelle (traitement du « BIG DATA »).

Bibiographie

AFZAL M.T., BALKE W.T., MAURER H., KULATHURAMAIYER N. Improving citation mining. 2009. Networked Digital Technologies, 2009. NDT'09. First International Conference on 15 (11) p. 116–121

COHEN William, RAVIKUMAR Pradeep, FIENBERG Stephen E., 2003 : A Comparison of String Distance Metrics for Name-Matching Tasks

CORTEZ Eli, da SILVA Altigran S., GONÇALVES, Marcos André, MESQUITA Filipe, de MOURA, Edleno S. FLUX-CIM: Flexible Unsupervised Extraction of Citation Metadata. 2007. Proceedings of the 2007 conference on Digital libraries - JCDL '07 p. 215

DAY Min-Yuh,TSAI Richard Tzong-Han, SUNG Cheng-Lung, LEE Cheng-Wei. A knowledge-based approach to citation extraction. Conf. 2005. IRI- p. 4-9

DAY Min-Yuh, TSAI Richard Tzong-Han, SUNG Cheng-Lung, HSIEH Chiu-Chen, LEE Cheng-Wei, WU Shih-Hung, WU Kun-Pin, Chomg-Shyong ONG, Wen-Lian HSU. Reference metadata extraction using a hierarchical knowledge representation framework. 2007. Decision support systems 43 (1) p. 152-167

GALIBERT, O., ROSSET, S., TANNIER, X., and GRANDRY, F. Hybrid Citation Extraction from Patents. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*. La Valette, Malta, May 2010. ELRA.

GALIBERT, O., ROSSET, S., GROUIN, C., ZWEIGENBAUM, P., & QUINTARD, L. (2011*). Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions*.

GILES C. LEE, BOLLACKER Kurt D, LAWRENCE Steve. Citeseer: an automatic citation indexing system. 1998. INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES p. 89 – 98

GROUIN, C., GALIBERT, O., ROSSET, S., QUINTARD, L., & ZWEIGENBAUM, P. (2010). *MESURES D'EVALUATION POUR ENTITES NOMMEES STRUCTUREES.*GROUIN, C. (2011). *MESURES UTILISEES DANS LES PROTOCOLES D'EVALUATION*. INALCO 2011-2012.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O., and QUINTARD L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc. of the Fifth Linguistic AnnotationWorkshop* (LAW-V), Portland, OR, june. Association for Com- putational Linguistics.

HETZNER Erik. A simple method for citation metadata extraction using hidden markov models. 2008. Proceedings of the 8th ACM/IEEE-CS joint conference p. 280

KIAT Ng Yong. Citation Parsing Using Maximum Entropy and Repairs Citation Parsing Using Maximum Entropy and Repairs. 2005

MAKHOUL J., KUBALA F., SCHWARTZ R., AND WEISCHEDEL R. (1999). PERFORMANCE MEASURES FOR INFORMATION EXTRACTION. IN *PROC. OF DARPA BROADCAST NEWS WORKSHOP*, PAGES 249–252.

NI Zhen. Automatic Citation Metadata Extraction Using Hidden Markov Models. 2009. Science and Engineering (ICISE), 2009 1st p. 802-805

PAUMIER, S. UNITEX 3.0Beta User Manual, Université Marne-la-Vallée

POWLEY Brett, DALE Robert. Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification. English p. 618-632

REITZIG, M. (2004). "The private values of 'thickets' and 'fences': towards an updated picture of the use of patents across industries." *Economics of Innovation and New Technology* 13(5): 457-476.