TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue

Florian Boudin
LINA - UMR CNRS 6241, Université de Nantes, France
florian.boudin@univ-nantes.fr

RÉSUMÉ

La recherche scientifique est un processus incrémental. La première étape à effectuer avant de débuter des travaux consiste à réaliser un état de l'art des méthodes existantes. La communauté francophone du Traitement Automatique de la Langue (TAL) produit de nombreuses publications scientifiques qui sont malheureusement dispersées sur différents sites et pour lesquelles aucune méta-donnée n'est disponible. Cet article présente la construction de *TALN Archives*, une archive numérique francophone des articles de recherche en TAL dont le but est d'offrir un accès simplifié aux différents travaux effectués dans notre domaine. Nous présentons également une analyse du réseau de collaboration construit à partir des méta-données que nous avons extraites et dévoilons l'identité du Kevin Bacon de *TALN Archives*, *i.e.* l'auteur le plus central dans le réseau de collaboration.

ABSTRACT

TALN Archives : a digital archive of French research articles in Natural Language Processing

Scientific research is an incremental process. Reviewing the literature is the first step to do before starting a new research project. The French Natural Language Processing (NLP) community produces numerous scientific publications which are scattered across different sources and for which no metadata is available. This paper presents the construction of *TALN Archives*, a digital archive of French research articles whose aim is to provide efficient access to articles in the NLP field. We also present an analysis of the collaboration network constructed from the metadata and disclose the identity of the Kevin Bacon of the *TALN Archives*, *i.e.* the most central author in the collaboration network.

MOTS-CLÉS : TALN Archives, archive numérique, articles scientifiques.

KEYWORDS: TALN Archives, digital archive, scientific articles.

1 Introduction

Mener des travaux de recherche scientifique de manière efficace suppose une analyse au préalable des travaux précédents du domaine. Cette étape d'analyse de la littérature existante permet d'évaluer la validité des idées proposées et d'identifier les contributions par rapport au domaine. L'avènement des moteurs de recherche a rendu cette tâche un peu plus simple, sans pour autant la

résoudre totalement. Parmi les difficultés qui subsistent et qui compliquent la tâche des moteurs de recherche, nous pouvons citer la dispersion des articles scientifiques sur les différents dépôts et bibliothèques numériques (e.g. HAL [1], Google Scholar [2], CiteSeer [3]), les erreurs de numérisation des versions papier des articles ou encore l'absence de méta-données associées pour l'indexation.

L'Association pour le Traitement Automatique des Langues (ATALA) organise annuellement les conférences TALN et sa session étudiante RÉCITAL. Ces dernières sont des événements majeurs pour la communauté francophone du Traitement Automatique de la Langue (TAL) et donnent lieu à de nombreuses publications scientifiques. L'ensemble des actes de chaque conférence est habituellement remis sur support physique aux participants et disponible sur le site web créé pour l'occasion. Ce mode de fonctionnement pose cependant plusieurs problèmes. Comment pérenniser l'accès aux actes ? Comment rechercher, parmi les différentes éditions de la conférence, les articles qui traitent d'une thématique particulière ? ceux écrits par un auteur particulier ? Le travail présenté dans cet article tente de répondre à ces questions en proposant la création d'une archive numérique francophone des articles scientifiques dans le domaine du TAL : *TALN Archives*.

Nos travaux s'inspirent de l'initiative menée par l'*Association for Computational Linguistics* (ACL) pour la construction de l'archive numérique *ACL Anthology* [4]. Créée en 2002, cette archive contient actuellement près de 22 000 articles scientifiques, pour la plupart rédigés en anglais, provenant de différents journaux, ateliers et conférences dans le domaine du TAL. Cette ressource offre un accès simple et rapide aux différents travaux de recherche menés depuis les quarante dernières années. L'*ACL Anthology* est en perpétuelle évolution et de nouvelles fonctionnalités y sont ajoutées régulièrement, comme récemment la recherche à facettes (Schäfer *et al.*, 2011) qui donne la possibilité aux utilisateurs de filtrer les articles selon différents critères tels que les énoncés (e.g. améliorer la qualité des traductions) ou les sujets abordés. Pour la construction de *TALN Archives*, nous souhaitons aller dans la même direction en portant toutefois une attention particulière aux méta-données qui seront utilisées pour l'indexation des articles.

Bien que l'utilisation première de l'*ACL Anthology* soit la recherche d'articles scientifiques, de nombreuses études l'ont utilisée comme corpus pour des tâches aussi variées que l'analyse de citations (Radev *et al.*, 2009), l'extraction d'information (Councill *et al.*, 2008), l'aide à l'écriture (Dale et Kilgarriff, 2010) ou l'analyse de sentiments (Athar, 2011). Dans cet article, nous présentons une analyse du réseau de collaboration construit à partir des méta-données et dévoilons l'identité du Kevin Bacon de *TALN Archives*, *i.e.* l'auteur le plus central dans le réseau de collaboration [5].

Dans la section 2, nous présentons la méthodologie de construction de l'archive numérique *TALN Archives*. En particulier, nous présentons les choix que nous avons fait concernant la structure de l'archive, le format de représentation des données et les méta-données associées aux articles. Dans la section 3, nous décrivons les expériences menées sur l'analyse du réseau de collaboration construit à partir des méta-données. Nous terminons cet article par une discussion sur les possibilités qu'offre *TALN Archives* et les travaux restants à effectuer.

1. http://hal.archives-ouvertes.fr
2. http://scholar.google.com
3. http://citeseerx.ist.psu.edu
4. http://aclweb.org/anthology-new/
5. http://fr.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon

2 Construction de *TALN Archives*

TALN Archives est une archive numérique dont le but à terme est de regrouper les articles scientifiques publiés dans le domaine du TAL par la communauté francophone. Notre objectif est d'offrir un portail unique permettant un accès pérenne et performant aux travaux effectués dans le domaine du TAL. Dans cet article, nous décrivons la première étape de ce travail, à savoir la construction des fichiers de méta-données à partir des articles scientifiques.

Chaque article scientifique dans *TALN Archives* possède un identifiant unique et un ensemble de méta-données (e.g. titre, auteurs, mots clés) que nous avons extrait à partir de son contenu. Ces dernières sont indispensables puisqu'elles sont utilisées, entre autres, par les moteurs de recherche pour indexer les articles. La visibilité et la diffusion des travaux de recherche sont donc fortement dépendantes de la qualité des méta-données. Elles sont également utilisées à des fins bibliographiques, avec par exemple la génération automatique de fichiers de références (e.g. BibTeX, EndNote), et pour l'interconnexion entre les différents dépôts et bibliothèques numériques.

Nous avons retenu le format XML pour le stockage des méta-données. La version de *TALN Archives* décrite dans cet article regroupe l'intégralité des actes des conférences TALN et RÉCITAL de 2007 à 2012, et contient 570 articles. L'ajout des actes des éditions plus anciennes (pré-2007), des articles de journaux (e.g. revue TAL [6]) et des actes publiés dans les différents ateliers associés aux conférences fait pour le moment partie des perspectives de ce travail.

2.1 Extraction des méta-données

Nous nous intéressons ici aux informations qu'il est possible d'extraire à partir des actes des conférences TALN et RÉCITAL. Deux types d'informations sont présents : celles associées aux éditions des conférences (e.g. dates, ville) et celles associées aux articles publiés (e.g. auteurs, titre, résumé). Pour *TALN Archives*, toutes les informations disponibles ont été extraites, avec le parti pris de ne pas se limiter aux méta-données issues du contenu des articles. Plusieurs autres informations, comme le nombre d'articles soumis ou les noms des présidents des comités de programme, ont été récupérées sur le site web de l'ATALA [7].

Les actes des conférences au format PDF ont été récupérés sur les sites web des conférences, ou à partir de support physique pour celles dont le site web n'est plus accessible [8]. Les articles ont ensuite été convertis au format texte à l'aide de l'outil PDFBox [9], puis nous avons étudié la possibilité de développer une méthode automatique pour l'extraction des méta-données.

Une des premières difficultés à laquelle nous nous sommes retrouvés confrontés concerne l'hétérogénéité des formats des articles. Les styles de soumission ont été largement modifiés au fil des années et un nombre important d'articles ne respectent pas les contraintes de style et de structuration pourtant imposées. L'application d'une méthode naïve, e.g. à base de patrons d'extraction, n'est donc pas envisageable. La conversion des fichiers PDF au format texte est également source de difficultés puisqu'elle supprime la structure du texte et introduit des erreurs

6. http://www.atala.org/-Revue-TAL-
7. http://www.atala.org/
8. Le site web de l'édition 2008 de TALN et de RÉCITAL n'est malheureusement plus accessible.
9. http://pdfbox.apache.org/

de césure et de segmentation. L'extraction automatique des méta-données n'est donc pas possible sans un travail important d'adaptation aux données et une certaine tolérance aux erreurs.

Comme nous souhaitions construire une ressource fiable, nous avons effectué l'extraction des méta-données de manière semi-automatique, avec une première étape automatique de pré-remplissage suivie d'une étape de correction et de complétion manuelle. Les données que nous avons construites ont été validées manuellement et pourront être utilisées dans le futur pour entraîner des outils d'extraction supervisées. La liste complète des méta-données que nous avons extraites est présentée ci-dessous. Les informations marquées d'un symbole * ont été récupérées à partir du site web de la conférence ou de celui de l'ATALA.

1. **Méta-données de la conférence**
   - Titre de la conférence, acronyme, ville, pays
   - Dates de début et de fin de la conférence*
   - Noms des présidents du comité de programme*
   - Formats des articles publiés (e.g. court, long)
   - Nombre d'articles soumis et nombre d'articles acceptés*
   - URL du site web de la conférence*
   - Identifiant(s) du(des) meilleur(s) article(s)*
2. **Méta-données pour chaque article**
   - Identifiant unique (e.g. taln-2008-long-001)
   - Noms des auteurs, emails, affiliations
   - Titre, résumé et mots clés (français et anglais si disponible)
   - Format de l'article
   - Numéros des pages
   - Nom de la session dans le programme

Contrairement à l'*ACL Anthology*, nous disposons, pour chaque article, d'un ensemble de mots-clés assignés par son(ses) auteur(s). La recherche des travaux portant sur une thématique particulière est donc grandement simplifiée. Il est en effet possible d'identifier des ensembles d'articles à partir d'un mot clé et ce même s'il n'apparaît pas dans le corps du document.

Nous notons également que 530 articles, parmi les 570 articles que compte l'archive, possèdent un résumé et des mots clés en français et en anglais. Cet ensemble de textes parallèles constitue une ressource intéressante pour des tâches comme la construction automatique de dictionnaires bilingues spécialisés (Fung, 1998) ou l'extraction de paraphrases (Barzilay et McKeown, 2001).

2.2 Statistiques de *TALN Archives*

La version de *TALN Archives* présentée dans cette étude est composée des actes des conférences TALN et RÉCITAL de 2007 à 2012. Au total, elle contient 570 articles scientifiques, 743 auteurs et 1 457 mots clés. Plus de 60 heures de travail ont été nécessaires pour vérifier et compléter les méta-données des articles. Les nombres d'articles publiés, d'auteurs ainsi que de mots clés pour chacune des éditions sont présentés dans la table 1.

Hormis pour les éditions 2008 et 2012, le nombre d'articles publiés est en constante augmentation, ce qui dénote une dynamique positive de la communauté francophone du TAL. Ces deux éditions coïncidaient avec l'organisation conjointe de TALN et des Journées d'Études sur la Parole (JEP).

	2007	2008	2009	2010	2011	2012	*TALN Archives*
# articles	88	66	104	106	117	89	570
# auteurs	163	128	246	220	231	186	743
# mots clés	335	242	329	341	335	316	1457

TABLE 1: Nombres d'articles publiés, d'auteurs et de mots clés pour les conférences TALN et RÉCITAL de 2007 à 2012.

Comme les travaux se situant à l'intersection des deux domaines ne peuvent être publiés dans les deux conférences, le nombre d'articles soumis à TALN a naturellement été plus faible (e.g. 104 soumissions pour TALN 2012, comparé aux 188 et 158 soumissions des éditions 2011 et 2010). Ce phénomène est d'ailleurs confirmé par un nombre plus restreint de mots clés, indiquant que les thématiques abordées dans les travaux sont moins nombreuses.

3 Analyse du réseau de collaboration

Les méta-données de *TALN Archives*, extraites à partir de chaque article scientifique, ont été utilisées pour construire un réseau de collaboration. Ce dernier est représenté sous la forme d'un graphe non dirigé $G = (V, E)$, où V est l'ensemble des nœuds et E l'ensemble des arêtes. Un nœud est ajouté au graphe pour chaque auteur dans *TALN Archives*. Lorsque deux auteurs ont collaboré sur un article, une arête est ajoutée entre leurs deux nœuds dans le graphe. Les poids des arêtes sont fixés en fonction du nombre d'articles auxquels les auteurs ont collaboré. Par exemple, l'arête entre les nœuds de deux auteurs ayant co-écrits trois articles aura un poids de trois.

La première expérience que nous avons menée porte sur l'identification des auteurs les plus centraux dans *TALN Archives*. Il s'agit d'identifier les auteurs qui ont un rôle majeur dans l'animation de la communauté francophone du TAL depuis les six dernières années. De nombreuses mesures ont été proposées pour calculer le degré de centralité d'un nœud dans un graphe [10]. Ici, nous utilisons la mesure de centralité harmonique (*Harmonic Centrality*) que nous calculons à l'aide de l'équation décrite ci-dessous :

$$C_{Harm}(V_i) = \sum_{V_j \in V,\ V_j \neq V_i} \frac{1}{\text{distance}(V_i, V_j)} \tag{1}$$

La table 2 présente la liste des dix auteurs les plus centraux dans *TALN Archives* selon la mesure de centralité harmonique. Pour chacun d'entre eux, nous reportons le nombre de collaborations (degré du nœud dans le graphe), le nombre d'articles ainsi que les mots clés apparaissant dans au moins deux de leurs articles. Il est intéressant de constater que cinq des dix auteurs les plus centraux travaillent sur la thématique de la traduction automatique statistique. Cet engouement pour la traduction automatique est également observable sur la totalité des actes contenus dans *TALN Archives* puisqu'il s'agit du mot clé le plus fréquent.

10. Une étude comparative des mesures de centralité dans un graphe est présentée dans `http://ecir2012.upf.edu/ecir_paolo_boldi.pdf`

Auteur	C_{Harm}	Deg.	Art.	Mots clés
Benoît Sagot	118.8	25	16	lexique-grammaire, résolution d'entités nommées, détection d'entités nommées, étiquetage morpho-syntaxique, lexique syntaxique, persan
Frédéric Béchet	115.8	15	7	reconnaissance automatique de la parole
Aurélien Max	111.8	12	11	paraphrase, wikipédia, aide à la rédaction, traduction automatique statistique
Delphine Bernhard	111.3	18	7	analyse syntaxique
François Yvon	110.9	14	8	traduction automatique statistique, alignement sous-phrastique
Karën Fort	107.3	11	7	accord inter-annotateurs, annotation manuelle, lexique
Philippe Langlais	106.7	9	10	traduction automatique statistique, analogie formelle, alignement de mots
Fabrice Lefèvre	106.0	9	5	compréhension de la parole, traduction automatique statistique, frames sémantiques, système de dialogue oral
Pierre Zweigenbaum	105.1	11	5	extraction d'information
Stéphane Huet	104.8	11	7	traduction automatique statistique, alignement de mots

TABLE 2: Liste des auteurs les plus centraux dans *TALN Archives*. Le nombre de collaborations (Deg.), le nombre d'articles (Art.) ainsi que les mots clés de fréquence supérieure à un sont également reportés.

La seconde expérience que nous présentons concerne l'identification du Kevin Bacon de *TALN Archives*, *i.e.* l'auteur le plus central dans le réseau de collaboration. Il s'agit d'identifier l'auteur qui possède la distance moyenne la plus faible avec les autres auteurs du réseau. Soit H le plus grand sous-graphe connecté de G, $|H|$ le nombre de nœuds dans le graphe H et distance(V_i, V_j) le plus court chemin entre les nœuds V_i et V_j, le nœud le plus central dans H est définit par :

$$\text{Centralité}(H) = \arg\min_{V_i \in H} \left[\frac{\sum_{V_j \in H,\ V_j \neq V_i} \text{distance}(V_i, V_j)}{|H| - 1} \right] \quad (2)$$

Dans *TALN Archives*, cet honneur revient à Frédéric Béchet qui possède une distance moyenne de 5,07 avec le reste des auteurs.

4 Discussion

Nous avons présenté la construction de *TALN Archives*, une archive numérique des articles scientifiques publiés dans le domaine du TAL par la communauté francophone. La version décrite dans cette étude est composée des actes des conférences TALN et RÉCITAL de 2007 à 2012. *TALN Archives* est dès à présent téléchargeable [11]. Une interface web permettant l'exploration et la

11. https ://github.com/boudinfl/taln-archives

recherche d'articles scientifiques dans *TALN Archives* est également disponible en ligne [12].

La première perspective à ce travail est bien entendu l'extension de l'archive aux actes des conférences TALN et RÉCITAL publiés avant 2007. Pour cela, les méta-données déjà contenues dans *TALN Archives* pourront être utilisées pour entraîner des méthodes d'extraction supervisées. La mise à jour de l'archive avec les actes des conférences futures est beaucoup plus simple puisque les outils de gestion de conférence couramment utilisés (e.g. *easychair*) permettent d'exporter les méta-données des articles acceptés.

Dans un second temps, nous souhaitons construire et analyser le réseau de citations de *TALN Archives*. Grace à ce dernier, il sera par exemple possible d'identifier les articles les plus influents pour une thématique donnée pour ensuite y extraire automatiquement les contributions principales. Des travaux récents ont également démontré l'utilité du réseau de citations pour améliorer la recherche d'articles scientifiques dans la littérature (Bethard et Jurafsky, 2010).

Remerciements

Nous tenons à remercier nos relecteurs anonymes pour leurs commentaires ainsi que les membres de l'équipe TALN du LINA pour leurs conseils avisés.

Références

ATHAR, A. (2011). Sentiment Analysis of Citations using Sentence Structure-Based Features. *In Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

BARZILAY, R. et MCKEOWN, K. R. (2001). Extracting Paraphrases from a Parallel Corpus. *In Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

BETHARD, S. et JURAFSKY, D. (2010). Who should I cite : learning literature search models from citation behavior. *In Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 609–618, New York, NY, USA. ACM.

COUNCILL, I., GILES, C. et KAN, M. (2008). ParsCit : An open-source CRF reference string parsing package. *In Proceedings of LREC*, volume 2008, pages 661–667. European Language Resources Association (ELRA).

DALE, R. et KILGARRIFF, A. (2010). Helping our own : text massaging for computational linguistics as a new shared task. *In Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 263–267, Stroudsburg, PA, USA. Association for Computational Linguistics.

FUNG, P. (1998). A statistical view on bilingual lexicon extraction : from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

RADEV, D., JOSEPH, M., GIBSON, B. et MUTHUKRISHNAN, P. (2009). A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Society for Information Science and Technology*, 1001:48109–1092.

12. http://www.florianboudin.org/taln_archives/

SCHÄFER, U., KIEFER, B., SPURK, C., STEFFEN, J. et WANG, R. (2011). The ACL Anthology Searchbench. *In Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 7–13, Portland, Oregon. Association for Computational Linguistics.