Un modèle d'annotation sémantique centré sur les utilisateurs de documents scientifiques: cas d'utilisation dans les études genre

Hélène de Ribaupierre and Gilles Falquet

ICLE, CUI, Université de Genève, Batelle, Genève, Suisse
helene.deribaupierre@unige.ch,Gilles.falquet@unige.ch

Résumé :
Lors de recherche de documents, les scientifiques ont des objectifs précis en tête. Nous avons mené des interviews auprès de scientifiques pour comprendre plus précisément comment ils recherchaient leurs informations et travaillaient avec les documents trouvés. Nous avons observé que les scientifiques recherchent leurs informations dans des éléments de discours précis, et non pas toujours dans le document en entier. A partir de cela, nous avons créé un modèle d'annotation prenant en compte ces éléments de discours. Nous avons implémenté ce modèle en OWL, et avons peuplé l'ontologie par des annotations provenant de documents dans le domaine des études genre. Nous montrons comment ce modèle permet de répondre à des requêtes précises et complexes sur un corpus de documents scientifiques.

Mots-clés : Ontologies, Annotation de documents scientifiques, Recherche d'information précise

1 Introduction

(Hannay, 2010) a écrit que les scientifiques ont de meilleurs outils pour gérer leurs données personnelles (photos et vidéo) que pour gérer ou chercher dans leurs données professionnelles. Ce constat est toujours valable, il est toujours difficile pour un scientifique de trouver le ou les bons documents qui correspondent effectivement à son besoin d'information. De plus, le nombre de documents scientifiques publiés chaque année est de plus en plus important (le nombre de documents dans Medline augmente de 0.5 millions par année (Novácek *et al.*, 2010)). Les moteurs de recherche académiques de type Google Scholar, DBLP ou Web of Knowledge indexent les documents par métadonnées et par les mots contenus dans le texte. Ils sont inefficaces dans le cas de requêtes complexes et précises telle que : « trouver tous les résultats de recherches qui utilisent une méthodologie quantitative et qui montrent que les filles sont meilleures dans les tâches de lecture que les garçons ». Pour traiter ce genre de requête, il faut entre autres être capable de détecter si les concepts cherchés apparaissent dans une partie du texte présentant un résultat de recherche ou une méthodologie. D'où le besoin d'indexer sémantiquement non seulement les mots des textes, mais également de caractériser la fonction de chaque fragment de texte (hypothèse, méthodologie, résultat, etc.). Dans cet article, nous proposons un modèle et un système d'annotation de documents scientifiques générique, prenant en compte les besoins des scientifiques. Nous avons choisi le domaine des études genre comme cas d'étude pour tester notre système, car les documents y sont très hétérogènes, allant d'études très empiriques à des documents de type philosophique, et n'utilisent que rarement le modèle structurel IMRaD (introduction, méthodologie, résultat et discussion).

2 Modèle d'utilisateur

Parmi les travaux qui étudient le comportement de recherche d'information et de lecture des scientifiques, (Bishop, 1999), en indexant des composants spécifiques dans une bibliothèque numérique (figures, conclusions, références, titres, titre de figures/tableaux, auteurs, etc.), montre que les scientifiques les utilisent pour faire des recherches d'information plus pertinentes. (Renear & Palmer, 2009), montrent que les scientifiques lisent et extraient des informations spécifiques telles que les "findings" [1], les équations, les protocoles de recherches et les données.

Pour comprendre ces besoins, nous avons mené deux études, l'une quantitative, l'autre qualitative, auprès de scientifiques de différentes communautés (de Ribaupierre & Falquet, 2011). Ces entretiens nous ont aussi permis de construire des cas d'utilisation génériques à partir des questions qu'ils se posaient avant d'utiliser un moteur de recherche et de convertir leurs questions en mots-clés. Nous avons extrait une douzaine de cas d'utilisation, dont trois exemples sont présentés ci-dessous.

Exemples de questions que les interviewés se posent	**Cas d'utilisation extrait**
Trouver les définitions de la notion d'homogénéité sémantique et si cela se calcule.	Trouver les différentes définitions d'un terme, et leurs différentes facettes [2]
Est-ce que Christine Delphy se dispute avec Patricia Roux dans un article ?	Trouver les auteurs qui ne sont pas d'accord avec l'auteur X, ou inversement, trouver les auteurs qui sont en accord avec l'auteur X
Trouver tous les auteurs qui travaillent sur la variabilité intra-individuelle du point de vue du comportement	Trouver les auteurs dans mon domaine de recherche

Nous avons aussi trouvé, que les scientifiques se concentrent sur certaines parties des documents qu'ils lisent. Les cinq types d'information que les scientifiques regardent en priorité sont (en dehors du résumé), les findings, les méthodologies, les hypothèses, les définitions et les travaux référencés (background). En entretien, nous avons confirmé l'hypothèse selon laquelle ces types ne correspondent pas forcément (même si souvent appelés de la même manière) aux parties structurelles des documents, mais bien à des fragments décrivant un de ces types pouvant se trouver n'importe où dans le document. Ainsi quand un scientifique parle de findings, il ne se réfère pas forcément à la partie structurelle portant le nom "findings", mais à tous les findings contenus dans le document.

3 Modèle d'annotation d'articles scientifiques

Il existe un certain nombre de modèles d'annotation pour les documents scientifiques. Certains auteurs (Groza *et al.*, 2007; Harmsze, 2000; Teufel & Moens, 2002; Ibekwe-Sanjuan *et al.*,

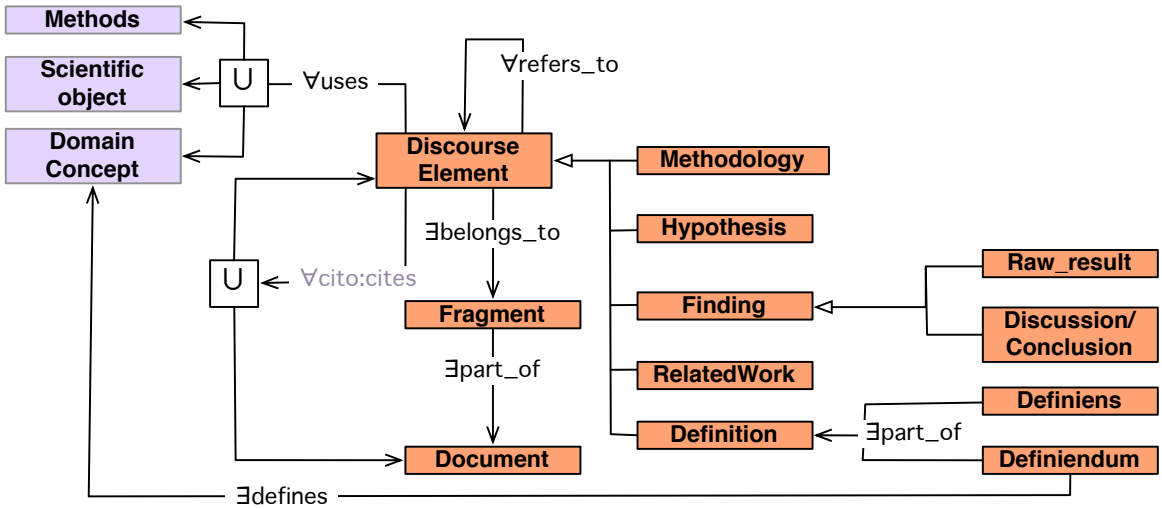
1. Il est difficile de trouver une traduction satisfaisante du mot "findings" qui regroupe à la fois les résultats bruts, les différentes trouvailles, observations discussions et conclusions du document scientifique

2. Notons à propos de cette requête que l'emploi de l'option "define" de Google est loin d'être satisfaisant. En effet, Google ira chercher des définitions provenant de glossaires connus ou de Wikipédia, donc "grand public" ou consensuelles, alors que le but de cette requête, pour le scientifique, est au contraire de trouver les définitions d'un terme proposées par des scientifiques dans un corpus de documents donné.

2008; de Waard *et al.*, 2009) proposent d'utiliser les structures rhétoriques ou les éléments de discours pour annoter les documents de manière à produire de meilleurs systèmes de recherche d'information ou pour créer des résumeurs automatiques. Ces travaux utilisent généralement des documents provenant des sciences dites "dures" telles que la biologie, la médecine ou la physique, où les documents sont très structurés, et où l'écriture des résultats, hypothèses ou méthodes est souvent formelle. Par ailleurs, seuls (Djioua & Descles, 2007), prennent en compte la définition comme type d'élément de discours. De plus, à notre connaissance, seul (Teufel & Moens, 2002) et (Djioua & Descles, 2007), annotent automatiquement les documents, les autres modèles sont utilisés pour de l'annotation manuelle. La construction de notre modèle d'annotation (voir figure 1) est essentiellement basée sur l'analyse des entretiens et du questionnaire, mais elle cherche aussi à agréger certains des modèles proposés.

Notre modèle d'annotation est basé sur trois axes : les éléments de discours, les relations explicites entre les documents et les métadonnées.

FIGURE 1 – Modèle d'annotation de document scientifique (Methods, Scientific object et Domain concept sont importés d'autres ontologies)



Eléments de discours : Ces éléments sont définis par l'ontologie SciDocAnnotation. Le contenu des éléments de discours est indexé sémantiquement à l'aide de concepts provenant d'ontologies auxiliaires : 1) ontologie(s) du domaine étudié ; 2) ontologie(s) des objets scientifiques (équations, modèles, algorithmes, théorème, etc.) ; 3) ontologie(s) des méthodes (types de méthodes, types de variables, outils utilisés, etc.). Il y a cinq types d'éléments de discours

Une Definition se compose d'un Definiens (la phrase définitoire) et d'un Definiendum (le terme défini). Le définiendum est relié par la relation defines à un concept du domaine, ainsi les différentes définitions d'un même définiendum utilisent le même concept du domaine.

Un Finding regroupe toutes les trouvailles, observations, discussions et conclusions d'un document. Le Raw_result qui défini des résultats pas encore analysés ou discutés. Alors que la Discussion/Conclusion, comme son nom l'indique, définit des résultats analysés et discutés.

Un élément de type Methodology décrit les différentes méthodes et étapes utilisées dans la recherche.

Un élément de type Hypothesis est une proposition de réponse à une question posée.

Un élément, quel que soit son type, est également de type RelatedWork s'il provient d'autres travaux. Nous sommes partis de l'hypothèse que les annotations des documents scientifiques doivent se faire sur une base de connaissance "universelle", et non pas centrée sur l'auteur. Nos études ont montré que ce ne sont pas forcément les écrits d'un auteur précis qui intéressent les scientifiques, mais une connaissance plus "universelle", qui doit par la suite être réattribuée à son auteur. Par exemple, quand une personne interrogée a répondu : « Trouver les différents articles qui traitent de l'évaluation des simulateurs chirurgicaux », cette scientifique, dans un premier temps, est intéressée à trouver tous les documents traitant de ce sujet, et cela indépendamment de l'auteur.

Références explicites d'un document/élément de discours à un autre document/élément de discours : Nous utilisons pour cela l'ontologie CiTO [3] de (Shotton, 2009). Contrairement aux moteurs de recherche académiques cités ci-dessus, nous annotons les relations, non pas au niveau document/document, mais au niveau élément de discours/élément de discours. Cela permet de résoudre des questions précises telle que « tous les findings démontrant une différence de sexe à l'école en mathématique et référençant un *résultat* de *Zazzo* ». Il devient également possible d'effectuer des analyses plus fines du réseau des citations, en fonction des types d'éléments cités ou citants.

Métadonnées : Il s'agit des données bibliographiques usuelles, telles que le nom des auteurs, le titre de l'article, le nom du journal ou de l'éditeur, etc.

4 Implémentation et Evaluation du modèle sur un cas dans le domaine des études genre

Nous n'avons pas trouvé d'ontologie du domaine des études genre, ni des objets scientifiques, ni des méthodologies que nous aurions pu importer dans notre modèle. Nous avons donc créé ces ontologies (contenant respectivement 465, 19 et 36 classes) [4].

Nous avons utilisé GATE [5], ANNIE [6], les règles JAPE et les modules de gestion d'ontologies, pour annoter automatiquement les différents éléments de discours contenus dans les documents et les concepts décrivant le contenu. Dans le but d'automatiser l'annotation des documents, nous avons défini l'élément de discours au niveau de la phrase, et le fragment au niveau du paragraphe. Après analyse manuelle d'un corpus de 20 documents, nous avons analysé les motifs syntaxiques des types d'éléments de discours à l'aide d'ANNIE et créé des règles JAPE (20 règles pour les findings, 34 règles pour les définitions, 11 règles pour les hypothèses et 19 règles pour les méthodologies).

Pour tester la qualité de l'annotation automatique nous avons annoté manuellement 555 phrases en études genre [7], créant ainsi un "golden standard". Nous avons effectué des mesures de précision/rappel sur ces phrases (voir tableau 1) qui montrent une bonne précision, mais un rappel peu élevé.

3. http ://purl.org/spar/cito

4. disponible sous http ://cui.unige.ch/~deribauh/Ontologies/

5. http ://gate.ac.uk/

6. http ://gate.ac.uk/ie/annie.html

7. les éléments de discours annotés manuellement, ne viennent pas des mêmes documents sur lesquels nous avons fait nos analyses syntaxiques.

TABLE 1 – Mesures de précision/rappel

Types d'éléments de discours	Nb de phrases	Prec.	Rappel	F1.0s
findings	168	0.82	0.39	0.53
hypothèses	104	0.62	0.29	0.39
définitions	111	0.80	0.32	0.46
méthodologies	172	0.83	0.46	0.59

Nous avons ensuite annoté automatiquement 903 articles en anglais, venant de différents journaux (étude genre et sociologie). Nous avons importé ces annotations dans un triplestore Allegrograph [8]. Nous sommes arrivés, après nettoyage, à 73'994 fragments (paragraphe) et 342'425 éléments de discours (phrases) se répartissant en : 304'747 qui n'ont aucun type, 15'449 findings, 11'813 méthodologies, 7'244 hypothèses, 3'172 définitions, parmi lesquels 2'780 travaux référencés dont 1'444 findings, 792 méthodologies, 351 hypothèses et 193 définitions. Nous avons annoté toutes les phrases sans type exportées de GATE par un élément *SentenceNotDefined*. Nous utilisons cet élément dans des heuristiques nous permettant de pallier le taux de rappel (voir ci-dessus). En effet, si un fragment contient des phrases non définies et plus de trois phrases de même type et uniquement de ce type, nous inférons que les phrases non définies sont du même type que les phrases définies. Avec ces règles, nous avons pu identifier 341 findings, 130 méthodologies, 29 hypothèses et 6 définitions supplémentaires.

Pour effectuer des tests comparatifs avec des utilisateurs nous avons défini deux interfaces de recherche sur cette base d'annotations : une interface de recherche classique par mots clés (avec un modèle de pondération TF*IDF) et une interface à facettes basée sur notre modèle (les facettes correspondant aux types d'éléments de discours).

5 Discussion / Conclusion

La principale contribution de cet article est la catégorisation et l'annotation automatique des éléments de discours basé sur un modèle d'annotation construit à partir d'interviews et de questionnaires soumis à des scientifiques. On peut considérer que le modèle est réaliste dans la mesure où une annotation automatique des documents est possible avec des outils classiques de traitement de la langue naturelle. Si le taux de rappel est bas, la précision des annotations est bonne, signifiant que si un utilisateur lance une requête dans notre système, il a une bonne probabilité de trouver une information précise, contrairement aux moteurs de recherche académiques actuels. De plus, les exemples de questions précises que les scientifiques nous ont fournis constituent une base de cas d'utilisation et de requêtes pour valider notre modèle et évaluer notre système.

Divers tests avec des utilisateurs sont en cours pour comparer notre modèle aux modèles de recherche par mots clés. Nous ne disposons pas encore d'une analyse complète de ces tests, mais les premières mesures tendent bien à montrer que notre modèle est nettement supérieur en précision. À titre d'exemple, la requête «Trouver les définitions qui contiennent le mot *gender*»,

8. http ://www.franz.com/agraph/allegrograph/

livre effectivement 143 définitions contenant le mot *gender*, alors qu'une recherche avec le mot clé "gender" fournit 13'210 éléments de discours, ce qui est de peu d'utilité.

Pour permettre l'automatisation, nous avons dû simplifier certaines relations du modèle de départ. Dans le cas des définitions, nous annotons les définitions, mais pas encore le définiendum des définitions. Une autre simplification que nous avons dû introduire concerne les relations entre documents, la granularité de la cible des citations est encore insuffisante (document au lieu de l'élément de discours).

Références

BISHOP A. P. (1999). Document structure and digital libraries : how researchers mobilize information in journal articles. *Information Processing and Management*, **35**(3), 255 – 279.

DE RIBAUPIERRE H. & FALQUET G. (2011). New trends for reading scientific documents. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, BooksOnline '11, p. 19–24, New York, NY, USA : ACM.

DE WAARD A., SHUM S. B., CARUSI A., PARK J., SAMWALD M. & SÁNDOR Á. (2009). Hypotheses, evidence and relationships : The hyper approach for representing scientific knowledge claims. In *Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science, Springer Verlag : Berlin*.

DJIOUA B. & DESCLES J. (2007). *Indexing documents by discourse and semantic contents from automatic annotations of texts*.

GROZA T., MULLER K., HANDSCHUH S., TRIF D. & DECKER S. (2007). Salt : Weaving the claim web. In *Proceedings of the Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea (Berlin, Heidelberg*.

HANNAY T. (2010). What can the web do for science ? *Computer*, **43**(11), 84–87.

HARMSZE F. (2000). *A modular structure for scientific articles in an electronic environment*. PhD thesis.

IBEKWE-SANJUAN F., SILVIA F., ERIC S. & ERIC C. (2008). Annotation of Scientific Summaries for Information Retrieval. In O. A. . H. ZARAGOZA, Ed., *ECIR'08 Workshop on : Exploiting Semantic Annotations for Information Retrieval*, p. 70–83, Glasgow, Royaume-Uni.

NOVÁCEK V., GROZA T., HANDSCHUH S. & DECKER S. (2010). Coraal - dive into publications, bathe in the knowledge. *J. Web Sem.*, **8**(2-3), 176–181.

RENEAR A. H. & PALMER C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing (vol 325, pg 828, 2009). *Science*, **326**(5950), 230–230.

SHOTTON D. (2009). Cito, the citation typing ontology, and its use for annotation of reference lists and visualization of citation networks. *The 12th Annual BioOntologies Meeting*, p. 1–4.

TEUFEL S. & MOENS M. (2002). Summarizing scientific articles : experiments with relevance and rhetorical status. *Computational linguistics 28*, **4**, 409–445.