HAL
open science

Ordonnancement des réponses pour une recherche d'information sémantique à partir d'une ontologie discursive

Iana Atanassova, Marc Bertin, Jean-Pierre Desclés

To cite this version:

Iana Atanassova, Marc Bertin, Jean-Pierre Desclés. Ordonnancement des réponses pour une recherche d'information sémantique à partir d'une ontologie discursive. Quatrième Atelier Recherche d'Information SEmantique RISE (RISE2012) co-located with 12e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, https://sites.google.com/site/frenchsemanticir/home/rise-2012, Jan 2012, Bordeaux, France. hal-01954027

HAL Id: hal-01954027

https://hal.science/hal-01954027v1

Submitted on 13 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Ordonnancement des réponses pour une recherche d'information sémantique à partir d'une ontologie discursive

Iana Atanassova, Marc Bertin, Jean-Pierre Desclés

LaLIC (Langues, Logiques, Informatique, Cognition)
Université Paris-Sorbonne
Maison de la recherche
28 rue Serpente 75006 Paris
{iana.atanassova | marc.bertin | jean-pierre.descles}@paris-sorbonne.fr
http://lalic.paris-sorbonne.fr

Résumé. Nous avons élaboré un système de recherche d'information sémantique en utilisant les annotations automatiques issues d'une ontologie discursive. Le système permet d'effectuer des recherches selon différentes catégories sémantiques liées aux citations bibliographiques dans des publications scientifiques. Nous avons proposé de nouveaux critères d'ordonnancement qui exploitent les annotations et la structure de lontologie. Ces critères prennent en compte la spécificité et la diversité des annotations, ainsi que les types de marqueurs linguistiques. Nous avons obtenu un score de pertinence que nous avons évalué en comparant les résultats du système avec des jugements de pertinence humains. L'évaluation montre que les algorithmes proposés produisent un ordonnancement de qualité pour les résultats se trouvant en début de liste.

1 Introduction

L'annotation sémantique permet d'accéder au contenu textuel de façon plus pertinente que par une recherche uniquement par mots clés. Il s'agit d'étendre le modèle de recherche classique afin de prendre en compte des annotations sémantiques automatiques des relations dans des textes (*hypothèse, rencontre, définition, citation,* ...), qui sont identifiées automatiquement par un moteur d'annotation. Notre approche de la recherche et extraction d'informations revient à croiser deux types de requêtes, l'une portant sur la sélection d'un point de vue et l'autre sur un ensemble de termes. Les points de vue correspondent aux catégories annotées dans les textes et qui permettent de filtrer l'ensemble des occurrences des termes de la requête. Si les annotations font partie de l'index du système, elles sont également utilisées par le langage des requêtes, donnant ainsi la possibilité d'exprimer le besoin informationnel à partir de relations sémantiques.

2 Ontologie linguistique et annotation sémantique automatique

Nous traitons des corpus annotés automatiquement par des catégories discursives, organisées dans une ontologie linguistique, appelée également *carte sémantique*. Une carte sémantique est le produit d'une conceptualisation des relations sémantiques dans les textes. Elle est un treillis dont les nuds sont des classes de concepts, et dont les arcs orientés représentent des liaisons de spécifications et de généralisations entre ces classes. Les concepts (ou points de vue) dans la carte sémantique peuvent être de nature grammaticale (par exemple *accompli/inaccompli*) ou discursive (par exemple *annonce thématique, rencontre, définition*). Les catégories sont instanciées par des classes de marqueurs linguistiques, appelés *indicateurs*, qui sont des expressions (continues ou discontinues) porteuses de la sémantique de chaque catégorie. Ces expressions sont identifiables à la surface des textes. Étant donné que les formes linguistiques sont souvent polysémiques, l'occurrence d'un indicateur dans un segment n'est pas suffisante pour attribuer la catégorie d'annotation. La désambiguïsation se fait par l'examen du contexte en vérifiant la présence ou absence d'un certain nombre d'indices contextuels. L'annotation s'inscrit ainsi dans la méthode d'Exploration Contextuelle (EC) Desclés et al. (1997) qui est une technique opératoire permettant de tenir compte du contexte pour lever l'indétermination sémantique des formes linguistiques.

Au-delà d'être un réseau de concepts, la carte sémantique organise les marqueurs et les règles d'EC sous-jacentes dans un réseau avec des relations de spécification, entre les concepts, ainsi qu'entre les classes de marqueurs linguistiques, et d'instanciation, entre chaque concept et la classe de marqueurs qui lui sont associés. Elle est issue d'une étude linguistique qui a pour but de systématiser les marqueurs linguistiques par lesquels points de vue se réalisent à la surface des textes.

Les règles d'EC sont écrites sous forme déclarative, où la vérifications de la présence ou de l'absence de formes de surface spécifiques (indices contextuels) déclenche l'annotation du segment (voir la figure 1). Comme le montre Atanassova (2012), la métode d'EC permet de reconnaître une classe de langages plus grande que celle des langages régulières et elle a une complexité linéaire par rapport à la longueur du segment annoté pour un ensemble de règles non-récursif.

Nous avons appliqué notre méthode en utilisant l'ontologie de la Bibliosémantique (voir la figure 2), qui propose une catégorisation des citations entre auteurs dans des publications scientifiques Bertin (2008, 2011). Elle est issue de l'analyse des contextes de références bibliographiques afin de dégager les différentes motivations d'un auteur pour citer une publication, par exemple pour introduire une méthode, comparer ses travaux aux autres, reprendre une définition, etc. L'annotation permet, à travers une étude de la bibliographie, d'identifier les relations entre auteurs. Les éléments de base pour l'annotation sont les phrases. Les figures 3 et 4 présentent quelques exemples de phrases annotées obtenues par le système. Les annotations multiples sont possibles lorsque la même phrase est porteuse de plusieurs relations sémantiques de l'ontologie.

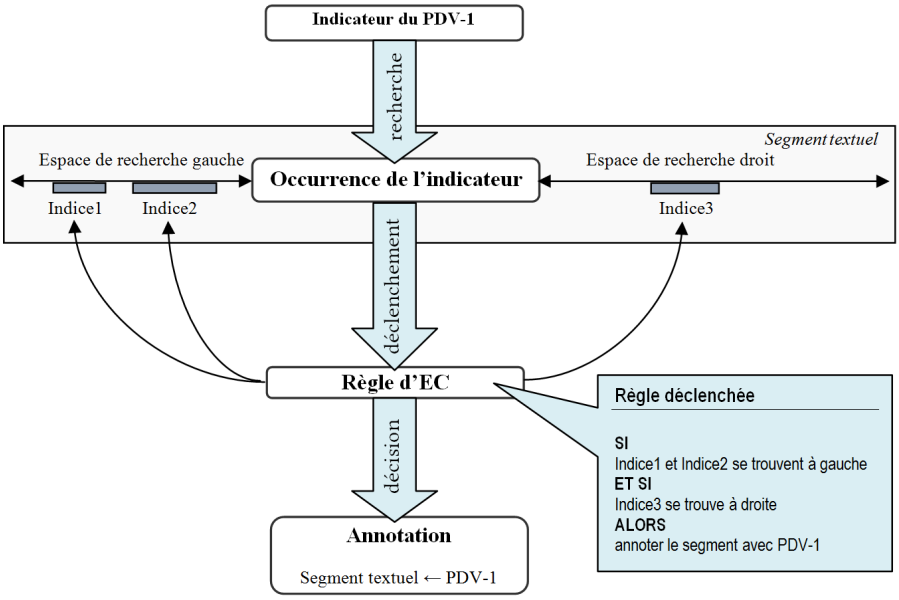

FIG. 1 – *Schéma de fonctionnement de la méthode d'EC*

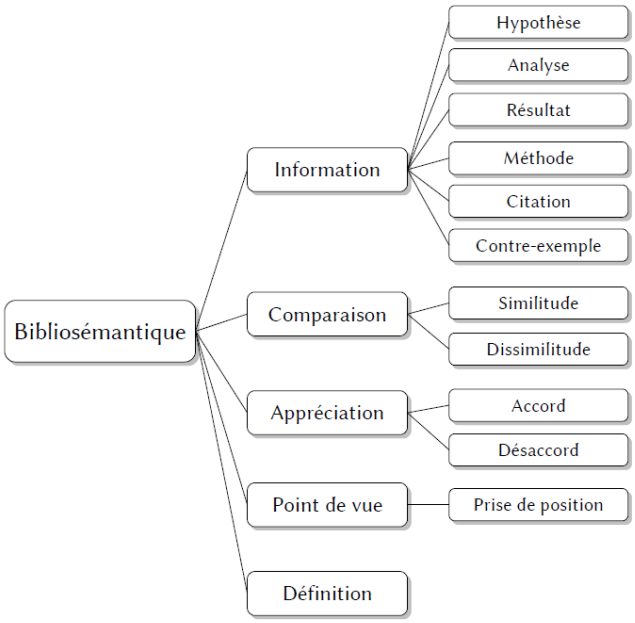

FIG. 2 – *Ontologie de la Bibliosémantique*

19. fr.utf8.lalic.these.Valliez.Chap3.txt.xml (These-Valliez)
A partir de ce principe, défini comme un « processus informatique de changements de représentations qui crée des représentations intermédiaires à différents niveaux » [DES 96, p. 105], fut proposée dans [Abraham, Desclés 92] une architecture cognitive d'interprétation sémantique des textes par des représentations iconiques (Fig. 111.2). (definition, résultat)

20. fr.utf8.lalic.these.pascu.Chap2.txt.xml (These-Pascu)
L'idée de définir le quantificateur comme un opérateur est reprise plus tard et indépendamment par Curry [Cur58] . (definition, similitude)

21. fr.utf8.alsic.2006.v9.ADEN.txt.xml (ALSIC2006)
Deyrich ([Deyrich05] : 127) rappelle la notion " d'illusion d'optique sociale" introduite par Martinon [Martinon96] par laquelle les sujets perçoivent les discours multimédiatiques à partir de leurs référents culturels, en effet, la mise en réseau des informations collectées reste, le plus souvent, ethnocentrée. (definition, citation)

FIG. 3 – *Phrases annotées : définition*

18. TheseIanaFinale.txttest.xml (Erudit)
Ces techniques introduisent un certain biais par rapport aux résultats de recherche et la visibilité des pages [Samier et al., 2007] : ce n'est plus le contenu d'une page qui est le principal facteur de son positionnement, mais l'expertise technique du créateur du site et l'argent investi dans la maintenance [Weideman, 2004]. (information, prise de position)
Fiche de synthèse

19. fr.utf8.intellectica.2002.Vol1.Num34_11_Lecuyer.txt.xml (Intellectica1991-2002)
Parfois même, c'est un paradigme expérimental précis qui est contesté (Haith, 1998; Cashon & Cohen (2000) . (prise de position)
Fiche de synthèse

20. fr.utf8.lalic.these.Bertin.Complete5.txt.xml (These-Marc)
Les principales critiques ont été exprimées par [Seglen, 1997] ,[Amin et Mabe, 2003], [Monastersky, 2005], [Ewing, 2006], [Adler, 2007], [Hall, 2007]. (prise de position)
Fiche de synthèse

FIG. 4 – *Phrases annotées : prise de position*

3 Ordonnancement des réponses

3.1 Méthodes existantes

La pertinence des résultats dans une grande partie des moteurs de recherche est modélisée par une valeur numérique qui se calcule dans un premier temps à partir des mesures de similarité entre le document et la requête. Les mesures et les méthodes de calcul dépendent du modèle de recherche d'information employé. L'estimation de la pertinence obtenue, constituant le *score de contenu* Baeza-Yates et Ribeiro-Neto (1999), est calculée en temps réel puisqu'elle est dépendante de la requête. Parmi les fonctions d'ordonnancement simples, nous pouvons citer la fréquence des termes, représentant la somme des nombres d'occurrences des termes de la requête dans le document, ainsi que la mesure *tf-idf*. En se basant sur la mesure *tf-idf*, Robertson et al. (1995); Cummins et ORiordan (2006) définissent la fonction *Okapi-BM25* dans le cadre du modèle probabiliste.

Les scores indépendants de la requête sont liés à la nécessité de développer des moteurs de recherche efficaces sur le Web. Ces nouveaux scores peuvent rendre compte de l'importance et de la fiabilité d'un document dans un corpus hyper-textuel. Plusieurs alternatives aux scores de contenu ont été proposées comme celles s'intéressant à la structure du graphe hypertexte et les relations entre les documents au sein de la collectionKleinberg (1999); Page et al. (1998);

Baeza-Yates et al. (2006); Krishnan et Raj (2006). Ces nouveaux scores, appelés également *scores d'importance*, sont calculés hors ligne et expriment une pertinence relative du document pour une requête quelconque. Deux types d'information sont alors utilisées : la structure du corpus, exprimée par le graphe hyper-texte, et certains aspects du comportement des internautes. Les scores *PageRank* Page et al. (1998) et *HITS* Kleinberg (1999) mesure la popularité des pages web à partir des hyperliens. Des essais ont été faits pour améliorer l'efficacité du score PageRank en prenant en compte de nouvelles informations. Par exemple, Haveliwala (2003) propose *Topic-sensitive PageRank* et Richardson et Domingos (2002) introduit une variante de PageRank dépendante de la requête dans un modèle probabiliste. Le score *TrustRank* Gyöngyi et al. (2004) est une estimation de la fiabilité des pages, afin de détecter le spam, calculé de façon semi-automatique à partir d'un petit ensemble de pages contrôlées manuellement. Le score de ces pages est propagé à travers le graphe par les hyper-liens. Afin de pallier certains biais introduits par la structure hyper-textuelle, d'autres scores comme BrowseRank Liu et al. (2008) et TrafficRank Tomlin (2003) prennent en compte l'historique de la navigation web et considèrent le temps et la fréquence des visites effectuées par les internautes.

Récemment la problématique de l'ordonnancement a été traitée par des techniques d'apprentissage automatique. La pertinence étant une notion complexe et multi-dimensionnelle Saracevic (1970, 1996); Denos (1997); Simonnot (2002), l'utilisation de ces méthodes est motivée par le fait que l'apprentissage pourrait théoriquement rendre compte de multiples critères d'ordonnancement, sans pour autant devoir analyser et modéliser tous les phénomènes autour de la pertinence.

Étant donnée une collection de documents, l'apprentissage automatique des ordonnancements suppose l'existence d'un corpus d'entraînement, qui consiste en un ensemble de requêtes et un ensemble de documents jugés pertinents pour chacune des requêtes ainsi que les valeurs du score de pertinence. La première phase d'entraînement consiste à créer une fonction d'ordonnancement qui a la propriété de produire exactement toutes les listes ordonnées du corpus d'entraînement. Dans une deuxième phase de test, cette fonction est utilisée afin d'ordonner des réponses de nouvelles requêtes. La mise en uvre des algorithmes d'apprentissage des ordonnancements est confrontée à une difficulté importante, qui est la définition des critères d'apprentissage pertinents pour l'optimisation des mesures d'évaluation. En effet, les mesures telles que la précision moyenne MAP et $nDCG$, sont difficiles à optimiser directement. Une première approche a été proposée par Cohen et al. (1999), qui considèrent l'ordre relatif des documents en utilisant une fonction binaire de préférence entre chaque couple de documents. Cette idée a été exploitée par d'autres travaux, qui optimisent le rang moyen des documents pertinents en minimisant les erreurs dans la fonction de préférence Freund et al. (2003); Cao et al. (2006); Usunier (2006). Les approches plus récentes se focalisent sur l'optimisation des erreurs en début de la liste, par exemple Volkovs et Zemel (2009). Nous pouvons également citer AquaLog Lopez et al. (2005, 2007) qui est un système de question/réponse reposant sur un moteur d'inférence et utilisant une ontologie.

3.2 Une nouvelle méthode s'appuyant sur une ontologie discursive

Les annotations sémantiques dans le corpus reflètent en partie le contenu sémantique des segments annotés. Ainsi, nous considérons que ces annotations portent des informations sur la pertinence des segments, surtout lorsque la recherche s'effectue selon des requêtes utilisant les catégories d'annotation Desclés et Djioua (2006); Atanassova et al. (2008).

L'unité textuelle de base dans cette approche est la phrase : tous les documents sont segmentés en phrases qui constituent un contexte minimal pour le traitement. Une phrase est constituée d'un contenu textuel (une suite de termes) et des caractéristiques sémantiques qui sont attribuées à une partie ou à la totalité de la phrase. De même, les requêtes sont constituées de deux éléments non obligatoires : une suite de termes, reliés éventuellement avec des opérateurs logiques, et un ensemble de catégories sémantiques sélectionnés à partir de l'ontologie.

Soit une ontologie qui s'exprime par un graphe $C = (V, A)$, où V est l'ensemble des nuds et A est l'ensemble des arcs. Nous pouvons alors formuler les deux hypothèses suivantes :

Hypothèse 1 : Une phrase $p = (T_p, R_p)$ est pertinente pour la requête $q = (T_q, R_q)$, si $d(T_q, T_p) < \alpha$ et $R_q \subset R_p$, dans le cas où l'ensemble R_q est non-vide,
où T_p est le texte du segment p, T_q est le texte du segment q, $R_p \subset V$ est l'ensemble des catégories d'annotation de p, $R_q \subset V$ est l'ensemble des catégories d'annotation de q, et d est une distance définie entre segments textuels.

Hypothèse 2 : Les pertinences relatives des phrases annotées sont déterminées en partie par les relations annotées.

La pertinence peut être aussi évaluée en termes de proximité entre les relations annotées et le besoin exprimé par la requête. L'ontologie sous-jacente définit une organisation hiérarchique entre les catégories, en utilisant des relations de spécification et généralisation. Nous proposons des critères d'ordonnancement des réponses utilisant cette structure, permettant ainsi de définir un score d'annotation tenant compte des types de relations exprimées dans les phrases.

3.2.1 Position des catégories dans la hiérarchie

L'organisation des catégories dans l'ontologie est liée à une organisation des marqueurs linguistiques sous-jacents. Considérons un couple de catégories $(pdv_1, pdv_2) \in A$, où pdv_2 est plus spécifique que pdv_1 dans l'ontologie. Nous pouvons formuler les observations suivantes :

- Si une phrase a été annotée par pdv_1 et non par pdv_2, cela signifie qu'elle contient les marqueurs linguistiques permettant d'attribuer l'annotation pdv_1. Comme l'annotation par pdv_2 n'a pas été déclenchée, la phrase ne contient pas assez de marqueurs linguistiques permettant d'affiner l'annotation.
- L'annotation par pdv_2 implique l'annotation par pdv_1, puisque $(pdv_1, pdv_2) \in A$, c'est-à-dire pdv_2 est une spécification de pdv_1. Cela signifie qu'une phrase annotée par pdv_2 contient suffisamment de marqueurs linguistiques, indicateurs désambiguïsés par des indices contextuels, permettant de lui attribuer cette annotation ainsi que l'annotation par pdv_1.

Nous pouvons énoncer l'hypothèse suivante :

Hypothèse 3 : Si $(pdv_1, pdv_2) \in A$, une phrase annotée par pdv_2 est plus pertinente qu'une phrase annotée par pdv_1.

Pour un segment annoté $p = (T_p, R_p)$ nous pouvons alors définir le score suivant :

$$SPos(p) = \frac{1}{M} . \max_{pdv \in R_p} \{N(pdv)\}, \tag{1}$$

où M est le niveau maximal de l'ontologie, c'est-à-dire le niveau contenant les catégories les plus spécifiques $M = \max_{pdv \in V}\{N(pdv)\}$, et la fonction $N : R_p \rightarrow \mathbb{N}$ associe à chaque catégorie son niveau dans la structure de l'ontologie, la racine ayant le niveau 1. Le score $SPos$ représente le niveau de spécificité de l'annotation sémantique d'une phrase.

$$N(pdv) = \begin{cases} 1, & \text{si} \quad \nexists pdv_1 \in V \text{ t.q. } (pdv_1, pdv) \in A \\ k+1, & \text{si} \quad \exists (pdv_1, pdv_2, \ldots, pdv_k) \in V^k \text{ t.q. } (pdv_k, pdv) \in A \\ & \quad \text{et } (pdv_{i-1}, pdv_i) \in A, i = 2, \ldots, k. \end{cases} \tag{2}$$

Le score $SPos$ prend des valeurs entre 0 et 1, où 1 correspond au niveau le plus spécifique de la carte sémantique.

3.2.2 Annotations multiples

L'annotation multiple d'un segment textuel indique la présence de plusieurs relations sémantiques, identifiées pendant la phase d'annotation. Nous parlerons d'annotations multiples uniquement dans le cas où il n'existe pas de relation de spécification/généralisation entre les catégories d'annotation. Leur annotation s'effectue de façon indépendante. Un segment qui est annoté plusieurs fois cumule l'ensemble des relations sémantiques annotées grâce à la présence des marqueurs linguistiques permettant d'identifier chacune de ces relations. Aussi, nous supposons que la pertinence d'une phrase annotée augmente avec le nombre d'annotations qu'elle porte.

Pour un segment $p = (T_p, R_p)$, nous pouvons définir l'ensemble de ses annotations indépendantes dans $C = (V, A)$ de la façon suivante :

$$S_p = \{pdv \mid (pdv \in R_p) \ \& \ (\nexists pdv'(pdv' \in R_p) \& ((pdv, pdv') \in A))\} \tag{3}$$

Nous pouvons formuler l'hypothèse suivante :

Hypothèse 4 : La pertinence d'une phrase p est proportionnelle à la taille de S_p.

Nous définissons le score lié aux annotations multiples de la manière suivante :

$$SNum(p) = \sum_{pdv \in S_p} \frac{1}{M} \{N(pdv)\} \tag{4}$$

Le score $SNum$ représente la diversité de l'annotation sémantique au sein de l'ontologie.

3.2.3 Types de marqueurs linguistiques

Par définition, l'annotation d'un segment textuel s'appuie sur plusieurs marqueurs linguistiques, puisque l'occurrence d'un simple marqueur peut être polysémique. La nécessité d'identifier un grand nombre de marqueurs dans un segment traduit le fait que chacun de ces marqueurs est polysémique et ne suffit par pour attribuer l'annotation. En effet, si une règle d'annotation nécessite un seul marqueur pour effectuer l'annotation, nous pouvons considérer qu'il s'agit d'un marqueur qui, en plus d'être porteur de la sémantique de la relation de l'annotation, est non polysémique. Nous appelons ce type de marqueurs *forts*. Au contraire, si la phrase ne contient pas de marqueur fort, son annotation s'appuie sur plusieurs marqueurs, chacun entre eux étant polysémique. Dans ce deuxième cas, nous pouvons considérer que la phrase est moins pertinente qu'une phrase contenant un marqueur fort, ce que nous formulons dans l'hypothèse suivante :

Hypothèse 5 : Le nombre de marqueurs linguistiques ayant déclenché l'annotation d'une phrase est inversement proportionnel à la pertinence de la phrase.

Nous définissons le score suivant :

$$SType(p) = \sum_{pdv \in R_p} \frac{1}{K(pdv) + 1}, \tag{5}$$

où la fonction $K : R_p \rightarrow \mathbb{N}$ donne le nombre d'indices contextuels qui ont été utilisés. Le score $SType$ prend des valeurs réelles entre 0 et 1 pour une annotation simple.

3.2.4 Score d'annotation

Nous constatons que les trois scores $SPos$, $SNum$ et $SType$ que nous avons définis utilisent en partie des propriétés communes des segments annotés, comme le nombre d'annotations différentes. Par conséquent, les trois valeurs attribuées à un segment donné ne sont pas entièrement indépendantes. Afin d'exprimer les relations entre ces scores, considérons un segment p annoté par N_p points de vue. Nous avons les inégalités suivantes, nous avons les inégalités suivantes :

$$\left| \begin{array}{l} 0 < SPos(p) \leq 1 \\ 0 < SPos(p) \leq SNum(p) \leq N_p \\ 0 < SType(p) \leq N_p \end{array} \right. \tag{6}$$

Nous définissons une fonction d'ordonnancement qui combine les trois scores définis ci-dessus. Cette fonction est définie de façon à ce que le résultat de l'ordonnancement soit identique à l'application consécutive des scores $SPos$, $SNum$, $SType$. Nous l'appellerons *score d'annotation* puisqu'il s'agit d'un score issu principalement de l'annotation sémantique et qui traduit donc le degré de pertinence des phrases selon les annotations par l'ontologie linguistique. L'adéquation du score d'annotation ainsi défini se confirme par l'évaluation (voir la section 5).

La fonction d'ordonnancement peut être donnée sous la forme :

$$SAnnot(p) = \alpha_1 SPos(p) + \alpha_2 SNum(p) + \alpha_3 SType(p) \tag{7}$$

Le vecteur $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ définit une pondération entre les trois scores $SPos$, $SNum$ et $SType$, et dans un cas particulier, permet de définir un ordre d'application de ces critères. Les coefficients $\alpha_i \in]0, 1]$ sont fixés de façon à ce que :

$$\left| \begin{array}{ll} (SPos(p_1) > SPos(p_2)) & \Rightarrow (SAnnot(p_1) > SAnnot(p_2)) \\ (SPos(p_1) = SPos(p_2)) \& (SNum(p_1) > SNum(p_2)) & \Rightarrow (SAnnot(p_1) > SAnnot(p_2)) \end{array} \right. \tag{8}$$

Des valeurs possibles pour les coefficients α_i sont : $\alpha = (1; 0,25; 0,04)$. Ce vecteur satisfait les conditions (8) pour l'ontologie que nous étudions.

Le score d'annotation est calculé hors ligne et reste indépendant de la requête. Ce dernier peut alors être considéré comme un score d'importance. Cependant, plusieurs propriétés le différencient des autres scores d'importance existants. Premièrement, il ne dépend ni de la structure globale du corpus, ni du comportement de l'ensemble des utilisateurs. Deuxièmement, le score est issu indirectement du contenu textuel en se basant sur une annotation qui explicite des relations sémantiques présentes dans des textes. Dans ce sens, il se rapproche plus aux

scores de contenu, qui expriment une proximité entre la requête et les documents selon leurs contenus. Pour le calcul de notre score, l'évaluation d'un document dépend uniquement de son contenu et non pas du reste du corpus. Troisièmement, le score d'annotation s'appuie sur des ressources externes : il exploite, d'une part, les annotations automatiques, et d'autre part, la structure de l'ontologie linguistique qui encode certaines connaissances linguistiques.

4 Système

Nous nous intéressons ainsi au système informatique, permettant l'indexation de documents textuels structurés au format DocBook et annotés par des catégories discursives, proposant des traitements plein texte et des fonctionnalités de gestion des contenus, d'extraction et de génération de documents secondaires dans une perspective de recherche d'information. Le schéma de la figure 5 présente les principales étapes du traitement automatique. Le moteur d'annotation sémantique Excom Djioua et al. (2006); Alrahabi et Desclés (2008) produit des fichiers annotés en format XML qui sont ensuite indexés à l'aide d'une base de données dédiée. Les segments textuels annotés sont indexés en plein texte.

L'annotation par la méthode d'Exploration Contextuelle s'appuie sur des ressources linguistiques (les marqueurs et l'ontologie linguistique) qui sont également utilisées pour structurer les interfaces d'accès à l'information, offrant à l'utilisateur la possibilité de formuler des requêtes selon catégories discursives. De plus, l'organisation de l'ontologie et une partie des marqueurs sont utilisés par les algorithmes d'ordonnancement des réponses que nous avons mis en place spécifiquement pour ce type de systèmes.

L'interface, élaborée en PHP/AJAX, offre des fonctionnalités de recherche d'information et de navigation textuelle en se basant sur les annotations sémantiques. L'utilisateur peut visualiser les résultats sous forme de phrases extraites, à partir desquelles il peut accéder aux contextes des phrases, aux documents sources, ou à des nouvelles représentations des documents telles que des extractions utilisant les annotations et des bibliographies enrichies permettant de retrouver les segments textuels liés à chaque référence.

5 Évaluation

Notre objectif est d'évaluer l'algorithme d'ordonnancement parmi les dix premiers résultats de chaque requête. Notons que cette évaluation n'a pas pour but d'évaluer les performances globales du système. Une évaluation complète serait plus coûteuse du fait qu'un très grand nombre de résultats doit être examiné manuellement. Seules les phrases qui figurent parmi les résultats des requêtes ont été évaluées, ce qui rend l'évaluation inadéquate pour estimer le rappel.

L'évaluation consiste en deux étapes :

1. Établir les pertinences de référence, correspondantes aux attentes des utilisateurs pour les résultats d'un ensemble de requêtes. Pour cela, nous avons utilisé des jugements de pertinence attribués manuellement aux résultats du système par trois juges humains.
2. Comparer les résultats du système avec les valeurs de pertinence de référence afin de calculer les valeurs $P@i$, $nDCG$, ainsi que la précision moyenne MAP.

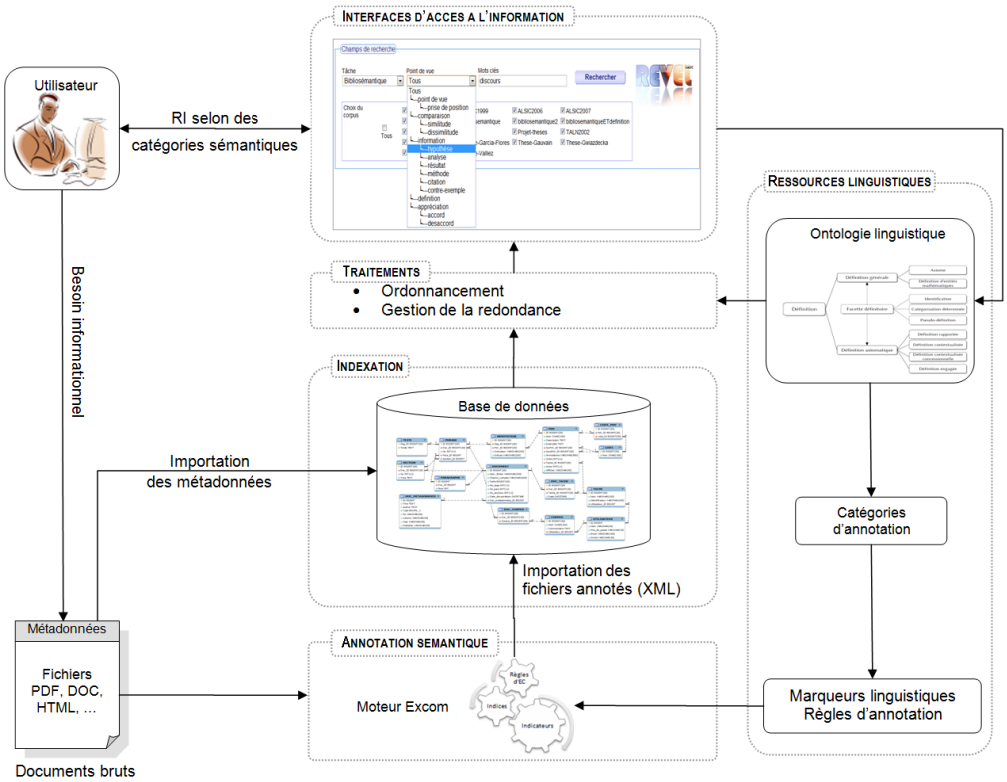

FIG. 5 – *Schéma général du traitement automatique*

5.1 Corpus

Pour tester le système nous avons utilisé un corpus de 330 documents de différentes tailles. Ce corpus comprend 8 thèses de l'Université Paris-Sorbonne, ainsi que des articles scientifiques issus de TALN 2002, Intellectica 1991-2002, ALSIC 1998, 1999, 2006 et 2007. La table 1 présente quelques statistiques du corpus traité. Tous les documents ont été annotés de façon automatique.

Phrases	Nb de mots	Mots par phrase	Phrases annotées	Pourcentage
359 354	7 890 558	21,96	8 672	2,41%

TAB. 1 – *Corpus*

Les annotations sémantiques représentent autour de 2,41% de tous les segments textuels. Cette annotation constitue un premier filtrage permettant d'identifier les phrases pertinentes pour l'analyse des citations bibliographiques. Pour la recherche d'information, nous considérerons que les réponses pertinentes se trouvent parmi les phrases annotées car elles sont porteuses des relations sémantiques recherchées, ce qui nous permet d'éliminer une grande partie du bruit parmi les résultats de la recherche d'information.

5.2 Protocole

Le corpus de requêtes pour l'évaluation consiste en 7 requêtes, où chaque requête est constituée d'une catégorie sémantique de recherche et un ou plusieurs mots clés. Pour ces requêtes, parmi toutes les phrases contenant les mots clés dans le corpus, le pourcentage des phrases annotées par les catégories recherchées varie entre 0,27% et 2,23%, ce qui montre que l'annotation permet dans un premier temps d'élire un ensemble très restreint de phrases pertinentes.

Les 10 premiers résultats de chaque requêtes ont été évalués par trois juges indépendants. Les jugements prennent trois valeurs possibles : 1 (*pertinent*), 0,5 (*peu pertinent*), ou 0 (*non pertinent*). Les résultats du système qui sont présentés aux juges sont ordonnés de façon aléatoire. Cet ordre est différent pour chacun des juges afin de s'assurer qu'il n'introduit pas de biais dans l'estimation de la pertinence. En dehors des étiquettes des annotations sémantiques et les contenus textuels des phrases, les juges n'ont aucune indication de l'importance ou de la qualité d'une phrase : ils n'ont pas accès aux valeurs des scores attribuées par le système, ni à l'ordonnancement qui en résulte.

Dans un premier temps, nous cherchons à estimer les valeurs de pertinence de référence, à partir des trois jugements pour chaque résultat. Pour un résultat r ayant obtenu les jugements $j_1(r)$, $j_2(r)$ et $j_3(r)$, nous considérons la moyenne :

$$Pert(r) = \frac{1}{3}\sum_{i=1}^{3} j_i(r) \tag{9}$$

Les valeurs $Pert(r)$ indiquent l'appréciation de la pertinence d'un résultat par l'ensemble des juges. Cette approche nous permet de modéliser la pertinence par une échelle à plusieurs valeurs entre 0 et 1, afin de mieux rendre compte du phénomène de pertinence. En effet, Kekäläinen et Järvelin (2002) soulignent que les jugements binaires sont à éviter pour les évaluations de la pertinence puisqu'ils ne peuvent refléter la variabilité et la complexité de celle-ci. Ils suggèrent que les jugements de pertinence soient plutôt modélisés par un ensemble de valeurs continu et proposent une généralisation des mesures de précision et de rappel pour de tels jugements.

Par la suite, étant donné un résultat r, nous considérerons que sa pertinence est représentée par la valeur $Pert(r) \in [0, 1]$. Nous allons considérer que r est non-pertinent si $Pert(r) = 0$.

5.3 Résultats

Pour chaque requête nous avons considéré l'ordonnancement des réponses par le système et nous avons calculé les valeurs $P@i$ ainsi que les valeurs $nDCG(i)$, où $i = 1, ..., 10$ est la position dans la liste produite par le système.

$$\begin{aligned} P@j = & \frac{1}{j}\sum_{i=1}^{j} Pert(r_i) \\ AP = & \frac{\sum_{j \in J} P@j}{|J|}, \text{ où } J = \{k \in [1; 10] | Pert(r_k) > 0\} \end{aligned} \tag{10}$$

Les précisions $P@j$ indiquent la précision en considérant la tranche des j premiers résultats, et AP donne une précision moyenne sur toute la liste ordonnée, en pénalisant d'avantage le bruit en début de la liste.

Nous avons également considéré la mesure $nDCG$ (*Normalized Discounted Cumulative Gain*) Järvelin et Kekäläinen (2002) qui utilise l'hypothèse que la pertinence diminue de façon logarithmique par rapport à la position dans la liste. La figure 6 donne les valeurs $P@i$ et $nDCG(i)$ obtenues à partir de l'ensemble des requêtes évaluées. L'abscisse représente le numéro de la réponse et l'ordonnée représente les valeurs pour les deux types de mesures.

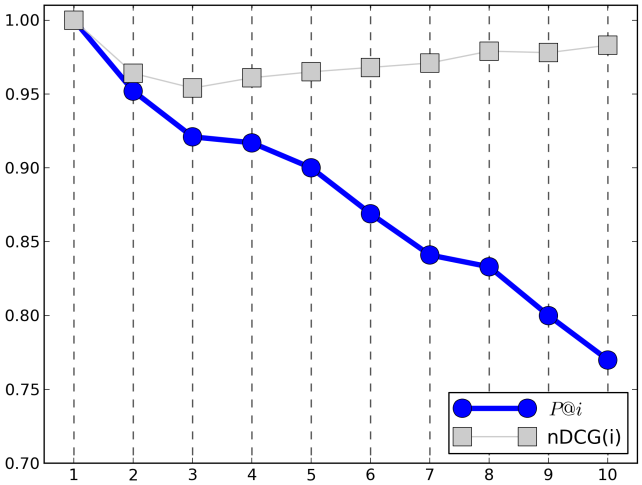

FIG. 6 – *Valeurs $P@i$ et $nDCG(i)$ pour les 10 permiers résultats*

Pour la précision moyenne MAP, nous avons obtenu :

$$MAP = \frac{1}{7}\sum_{i=1}^{7} AP(q_i) = 0,893. \tag{11}$$

6 Conclusion

Nous avons proposé un nouveau algorithme d'ordonnancement en nous appuyant sur la structure d'une ontologie discursive et sur l'annotation sémantique. L'évaluation montre que cet algorithme est très performant en ce qui concerne les premiers résultats fournis par le système. En effet, la figure 6 montre bien que le système a une tendance prononcée à classer les résultats les plus pertinents en début de la liste, et donc d'ordonnancer les résultats dans un ordre proche de celui des juges humains.

Cette première évaluation est encourageante, même si nous sommes conscients qu'il sera nécessaire de confirmer ce résultat à une plus grande échelle. Les valeurs obtenues pour la précision moyenne, notamment $MAP = 0,893$, confirment que l'annotation sémantique apporte une meilleure qualité à la recherche d'information. Au-delà de l'algorithme d'ordonnance-

ment, les valeurs des précisions très élevées reflètent également la précision des annotations sémantiques.

Références

Alrahabi, M. et J.-P. Desclés (2008). Automatic annotation of direct reported speech in Arabic and French, according to a semantic map of ennunciative modalities. In *6th Internatilnal Conference of NLP, GOTAL*, Gothenburg, Sweden.

Atanassova, I. (2012). *Exploitation informatique des annotations sémantiques automatiques d'Excom pour la recherche d'informations et la navigation*. Ph. D. thesis, Université Paris-Sorbonne.

Atanassova, I., J.-P. Desclés, A. Franchi, et F. Le Priol (2008). La plate-forme excom comme outil automatique d'annotations sémantiques des textes pour la catégorisation des informations sur le web. In *Colloque ń Internet : besoin de communiquer autrement ż, Université St. Clément d'Ohride*, Sofia, Bulgarie.

Baeza-Yates, R., P. Boldi, et C. Castillo (2006). Generalizing PageRank : damping functions for link-based ranking algorithms. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval 1*, 308–315.

Baeza-Yates, R. et B. Ribeiro-Neto (1999). *Modern Information Retrieval* (1e ed.). Addison Wesley.

Bertin, M. (2008). Categorizations and annotations of citation in research evaluation. In *FLAIRS*, Coconut Grove, Floride. AAAI Press.

Bertin, M. (2011). *Bibliosématique : une technique linguistique et informatique par exploration contextuelle*. Ph. D. thesis, Université Paris-Sorbonne.

Cao, Y., J. Xu, T. Liu, H. Li, Y. Huang, et H. Hon (2006). Adapting ranking SVM to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in IR*, pp. 186–193.

Cohen, W., R. Schapire, et Y. Singer (1999). Learning to order things. *Journal of Artificial Intelligence Research 10*, 243–270.

Cummins, R. et C. ORiordan (2006). Evolving local and global weighting schemes in information retrieval. *Information Retrieval 9*(3), 311–330.

Denos, N. (1997). *Modélisation de la pertinence en recherche d'information : modèle conceptuel, formalisation et application*. Ph. D. thesis, Université de Grenoble 1.

Desclés, J.-P., E. Cartier, A. Jackiewicz, et J.-L. Minel (1997). Textual Processing and Contextual Exploration Method. *CONTEXT'97, Rio de Janeiro 1*, 189–197.

Desclés, J.-P. et B. Djioua (2006). Machines d'annotation et d'indexation discursives de textes : EXCOM/MOCXE. In *Annotation automatique de relations sémantiques et recherche d'informations : vers de nouveaux accès aux savoirs*, Paris.

Djioua, B., J. G. Flores, A. Blais, J.-P. Desclés, G. Guibert, A. Jackiewicz, F. Le Priol, L. Nait-Baha, et B. Sauzay (2006). Excom : an automatic annotation engine for semantic information. *The 19th international FLAIRS Conference, Melbourne, Floride 1*, 285–290.

Freund, Y., R. Iyer, R. Schapire, et Y. Singer (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research 4*, 933–969.

Gyöngyi, Z., H. Garcia-Molina, et J. Pedersen (2004). Combating web spam with trustrank. In *Proceedings of the 30th international conference on Very large data bases*, Volume 30, pp. 576–587.

Haveliwala, T. (2003). Topic-sensitive pagerank : A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 784–796.

Järvelin, K. et J. Kekäläinen (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS) 20*(4), 422–446.

Kekäläinen, J. et K. Järvelin (2002). Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology 53*(13), 1120–1129.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*(5), 604–632.

Krishnan, V. et R. Raj (2006). Web spam detection with anti-trust rank. *2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Liu, Y., B. Gao, T. Liu, Y. Zhang, Z. Ma, S. He, et H. Li (2008). Browserank : letting web users vote for page importance. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 451–458.

Lopez, V., M. Pasin, et E. Motta (2005). Aqualog : An ontology-portable question answering system for the semantic web. *The Semantic Web : Research and Applications*, 135–166.

Lopez, V., V. Uren, E. Motta, et M. Pasin (2007). Aqualog : An ontology-driven question answering system for organizational semantic intranets. *Web Semantics : Science, Services and Agents on the World Wide Web 5*(2), 72–105.

Page, L., S. Brin, R. Motwani, et T. Winograd (1998). The pagerank citation ranking : Bringing order to the web. *Stanford InfoLab*.

Richardson, M. et P. Domingos (2002). The intelligent surfer : Probabilistic combination of link and content information in pagerank. *Advances in neural information processing systems 2*, 1441–1448.

Robertson, S., S. Walker, S. Jones, M. Hancock-Beaulieu, et M. Gatford (1995). Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126. NIST.

Saracevic, T. (1970). The concept of "relevance" in information science : A historical review. *Introduction to information science*, 111–151.

Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the 2nd Conference on Conceptions of Library and Information Science*, pp. 201–218.

Simonnot, B. (2002). De la pertinence à l'utilité en recherche d'information : le cas du Web. In *Recherches récentes en sciences de l'information : convergences et dynamiques, Actes du colloque Mics-Lerass*, pp. 393–410.

Tomlin, J. (2003). A new paradigm for ranking pages on the world wide web. In *Proceedings of the 12th international conference on World Wide Web*, pp. 350–355. ACM.

Usunier, N. (2006). *Apprentissage de fonctions d'ordonnancement : une étude théorique de*

la réduction à la classification et deux applications à la Recherche d'Information. Ph. D. thesis, Université Paris-VI.

Volkovs, M. et R. Zemel (2009). Boltzrank : learning to maximize expected ranking gain. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1089–1096. ACM.

Summary

We have developed an information retrieval system based on the automatic annotation using a discursive ontology. The system makes possible information retrieval of specific semantic categories related to bibliographic citations in scientific publications. We propose new ranking criteria which exploit the annotation and the structure of the ontology. These criteria take into account the specificity and the diversity of the annotations, as well as the different types of linguistic markers. We have defined a ranking function and we have evaluated it by a comparison between the system output and human relevance judgments. This evaluation confirms the algorithm efficiency for the ranking of highly relevant results.