Filtrar-S: nouveaux développements

Nicolas CAMPION[1] , Jacques CLOSSON[1],

Olivier FERRET[2], Romaric BESANÇON[2], Wei WANG[2],

Jin SHIN[3], Jean-Marc FLORET[3],

Brigitte GRAU[4], Xavier TANNIER[4],

Amar-Djalil MEZAOUR[5], Jean Marc LAZARD[5],

[1]ERAKLE, 42 rue d'Artois, 75008 Paris

[2]CEA LIST, 18 route du Panorama, BP 6, 92265 Fontenay-aux-Roses

[3] EXAKIS, 37-41, rue Louise Weills, 75013 Paris

[4]LIMSI, UPR-3251 CNRS-DR4, Bat. 508, BP 133, 91403 Orsay Cedex

[5]EXALEAD, 10 place de la Madeleine, 75008 Paris

nicolas_campion@yahoo.fr, jacques.closson@wanadoo.fr
olivier.ferret@cea.fr, romaric.besancon@cea.fr, wei.wang@cea.fr,
jsn@exakis.com, Jean-Marc.FLORET@exakis.com,
Brigitte.Grau@limsi.fr, Xavier.Tannier@limsi.fr
mezaour@exalead.com, jean-marc.lazard@exalead.com

Résumé –
Financé par l'ANR et administré par l'UTT avec pour objectif de contribuer à la sécurité du citoyen, le projet FILTRAR-S consiste à développer et à tester le démonstrateur d'un outil d'analyse automatique du contenu sémantique des textes écrits. L'outil utilise les statistiques d'occurrence des mots dans les documents textuels d'un corpus pour découvrir une classification thématique pertinente de leur contenu, et pour extraire des relations entre entités nommées qui sont présentes dans les documents. Du fait du caractère essentiellement non supervisé des méthodes utilisées, les catégories thématiques et les relations extraites des corpus n'on pas à être connues a priori. En outre, l'outil intègre un module d'indexation des topics et des relations découvertes dans les documents et un module de question/réponse pour la fouille du contenu des documents indexés. Cet article rappelle les principaux concepts et les objectifs du projet, présente les travaux effectués et les résultats obtenus à la fin de la deuxième année du projet, ainsi que les développements futurs.

Abstract – The project named FILTRAR-S is supported by ANR and managed by UTT with the aim of improving the citizens safety. It consists in the programming and testing of a tool that automatically analyzes the semantic content of any kind of written texts. The tool computes statistics of word occurrences in a corpora and discovers a relevant set of semantic topics that are instantiated in the texts of the corpus. The topics are then used to classify the texts and to extract the relations between pairs of entities names in texts. Because of the unsupervised nature of the methods used, the set of topic categories and the extracted relations does not have to be previously known and can be specific to a corpus. Then, a module indexes the texts through the topics and relations that have been discovered in their contents. At last, a module relies on the built index to find relevant answers to questions typed by users. This article points out the main theoretical bases and goals of the project, describes the work done and the results obtained at the end of the second year of the project. It also mentions the future developments and user tests that are on the agenda.

1. Problématique du projet

Le besoin de contrôler et d'exploiter la quantité croissante d'informations textuelles écrites qui se trouve stockée dans les messageries électroniques, les pages Web, les blogs ou les flux RSS, est devenu un problème majeur de nos sociétés, comme en témoigne l'actualité récente autour de Wikileaks.

Le contrôle et l'exploitation rapide de cette information qui transite sous la forme de documents contenant des textes écrits numérisés, et qui est souvent délivrée sous la forme de larges corpus, est fondamentale pour la lutte contre la cybercriminalité, le dépouillement de fichiers informatiques dans le cadre d'enquête judiciaires, la connaissance de la rumeur publique, la détection et la classification des nouveautés d'un domaine scientifique ou commercial etc.

La première caractéristique de FILTRAR-S est de s'intéresser au contenu sémantique de l'information textuelle véhiculée par les documents. Ainsi, contrairement à beaucoup de méthodes utilisées, les recherches d'information effectuées ne dépendent pas d'une correspondance exacte entre les mots des documents et ceux des structures de connaissance utilisées.

Mais c'est la seconde caractéristique qui fait l'intérêt et l'originalité de l'approche adoptée par FILTRAR-S: pour l'essentiel, les méthodes utilisées sont non supervisées, c'est-à-dire que le système construit lui-même, en les extrayant du type de corpus qu'il a à traiter, les thèmes sémantiques et les classes de relations qui vont constituer sa base de connaissances.

Un dernier aspect important du projet FILTRAR-S est que les méthodes non supervisées qu'il utilise constituent une modélisation plausible du fonctionnement de la mémoire sémantique. Ces méthodes exploitent la fréquence de co-occurrence des mots dans les documents textuels et, selon certains modèles influents, les traitements sémantiques contraints par ces co-occurrences sont responsables de l'acquisition et de la mise en oeuvre de processus associatifs qui caractérisent la mémoire sémantique humaine et permettent la récupération automatique de connaissances.

2. Architecture de FILTRAR-S

Le point d'entrée du démonstrateur FILTRAR-S est constitué par un module de crawling et de formatage des documents récupérés sur internet ou toute autre base de documents écrits numérisés. Les documents formatés font ensuite l'objet de traitements linguistiques visant éliminer les mots vides (grammaticaux) et à lemmatiser les autres mots. Ces traitements linguistiques ne sont qu'optionnels pour le module filtrage vers lequel les documents sont ensuite dirigés. Ce module filtrage procède d'abord à l'extraction non supervisée d'une série de topics ou thèmes développés dans les documents de larges corpus et il indexe les documents de ces corpus en fonction des topics qu'ils développent. Un filtrage binaire par acceptation/rejet des documents peut alors être effectué en calculant la similarité entre les topics associés à ces documents et les topics associés à une série de documents modèles, pré-sélectionnés dans une base sémantique. Ensuite les documents filtrés sont traités par le module d'extraction non supervisée de relations entre couples d'entités nommées. Ces entités sont identifiées par des traitements linguistiques supplémentaires et ils doivent être présents dans une même phrase. Le module identifie également les regroupements de relations équivalentes, ainsi que les regroupements de relations se raccrochant à un même thème. Comme pour les topics, les relations extraites sont associées aux documents sous la forme d'indexations et conservées avec ceux-ci dans une base de documents.

La fouille du contenu de la base de documents s'effectue par le biais d'un moteur de recherche ou par le biais d'un système de question-réponse. Dans le cas d'une fouille « orientée relation », la recherche se focalise plus exclusivement sur les relations indexées.

La figure 1 présente un schéma de l'architecture de Filtrar-S.

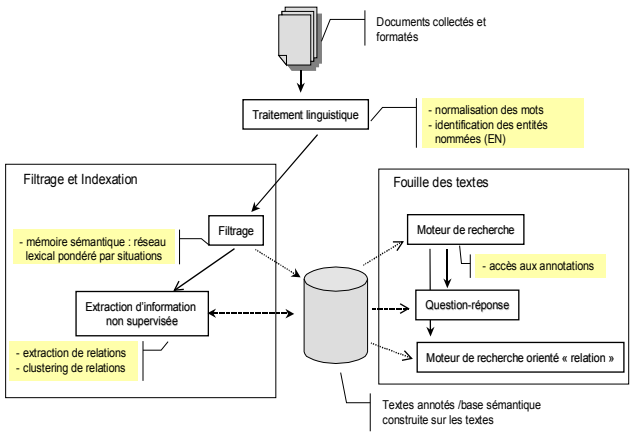

FIG. 1 : schéma de l'architecture Filtrar-S

3. Le module de filtrage de documents

3.1 Les principes de l'approche

L'objectif du module est de réaliser un filtrage sémantique des informations contenues dans les

documents. Il s'agit donc d'opérer une analyse de la signification des documents et non pas d'utiliser une correspondance terme à terme avec une série de mots clés.

Pour cela, notre approche consiste à simuler le fonctionnement associatif de la mémoire sémantique humaine. Les sciences cognitives ont largement montré que l'interprétation d'un document fait appel à nos connaissances des situations et des événements qui se déroulent habituellement dans le monde. D'autre part, des études expérimentales ont montré que le traitement sémantique d'un mot n'active pas seulement la signification de ce mot dans la mémoire des individus. Il y active aussi la signification des mots auxquels le mot traité est associé par le biais de connaissances sur le monde [2, 3, 4, 6]. Des travaux complémentaires ont montré que ces processus associatifs interviennent pendant la compréhension d'un document textuel. La convergence de ces associations vers un même thème permet la récupération des connaissances du monde correspondantes avec pour conséquence la sélection des significations contextuellement appropriées [16, 17] et la production d'inférences [1, 18].

Mais le problème qui a longtemps empêché toute modélisation opérationnelle de ces mécanismes est la nécessité d'introduire dans le système les grandes quantités de connaissances du monde qui sous-tendent les associations sémantiques entre mots et qui sont récupérées en retour, via le traitement des mots associés. Les choses ont changé avec l'arrivée des modèles statistiques exploitant la fréquence de co-occurrence des mots dans de larges corpus de documents. Leur force est d'exploiter la grande quantité de données réelles que constituent les comportements verbaux inscrits dans les documents et d'extraire ainsi des connaissances sous la forme de structures lexicales associatives. Certains de ces modèles ont été utilisés comme modèles du fonctionnement associatif de la mémoire sémantique et c'est sur l'un de ces modèles que nous nous appuyons pour la conception du module de filtrage.

3.2 Modèles statistiques permettant l'extraction de structures sémantiques associatives

À la suite du modèle vectoriel de Salton [7], une série de modèles ont exploité le comptage des cooccurrences des mots dans les corpus de documents pour permettre la récupération automatique de documents. Ceux qui intéressent le système FILTRAR-S sont ceux qui cherchent également à construire des structures sémantiques et modéliser le fonctionnement de la mémoire sémantique humaine [5]. En effet, simuler ce fonctionnement cognitif humain entraîne un parallélisme entre les inférences associatives humaines et les réponses du système, ce qui tend à résoudre les problèmes de pertinence et d'adaptation à l'utilisateur.

Le plus connu des modèles qui visent la simulation du fonctionnement cognitif est LSA ou Latent Semantic Analysis [8, 14]. Ce modèle utilise une technique mathématique qui réduit le nombre de dimensions d'une matrice Mot X Document, ce qui a pour effet de sélectionner les co-occurrences les plus significatives du corpus et d'extraire ainsi des structures sémantiques latentes qui contraignent les associations entre les mots représentés. Ce modèle donne des résultats qui ont une certaine validité psychologique, mais le principal problème est que les dimensions de l'espace ne sont pas directement interprétables. Il est donc difficile de sélectionner des sous-espaces vectoriels dans lesquels les nombreux mots polysémiques de tout corpus prennent un sens plutôt qu'un autre.

D'autres modèles ont fait l'économie de cette phase de réduction du nombre de dimension de la matrice, qui n'est en fait qu'une heuristique sans légitimité psychologique. Dans le modèle HAL, Burgess et Lund [9] comptabilisent les co-occurrences locales entre les mots d'un document au moyen d'une « fenêtre glissante » de 10 mots. Ils construisent par ce moyen une matrice mot X mot de très grande taille (70 000 X 70 000) et un espace vectoriel à 140 000 dimensions (car ils font aussi intervenir l'ordre dans lequel co-occurrent les mots). Comptabiliser ainsi les co-occurrences dans une fenêtre de taille limitée respecte le fait que les capacités de la mémoire de travail humaine limitent le nombre de mots dont les représentations lexicales (signifiants et signifiés) peuvent être maintenues simultanément actives. Cependant, à l'inverse de LSA, le modèle accorde une grande importance aux co-occurrences locales et il exige de grandes capacités de mémoire pour gérer le grand nombre de dimensions de l'espace.

Une autre approche est celle du random indexing que Jones & Mewhort [10] utilisent dans le cadre de leur modèle Beagle. Trois vecteurs sont utilisés pour représenter les mots dans un espace vectoriel dont le nombre de dimensions est fixé à priori. Un premier vecteur est censé coder la « forme perceptive » du mot et en l'état actuel du modèle, ses valeurs sont fixées aléatoirement. Par ailleurs, les valeurs de ce premier vecteur sont fixées de façon définitive. Un second vecteur est un vecteur « contexte » qui est recalculé à chaque occurrence du mot en faisant la somme des vecteurs perceptifs de chacun des mots qui co-occurrent avec ce mot. Sont comptabilisés comme co-occurrents tous les mots qui apparaissent entre deux signes de ponctuation successifs. Le troisième vecteur est un vecteur position qui comptabilise la distance moyenne avec laquelle chaque mot entre en co-occurrence avec ses voisins. Séduisant par son apprentissage contextuel cumulatif, ce modèle donne pour l'instant des performances équivalentes à celles du modèle LSA.

Le modèle dont s'inspire finalement le module de filtrage de FILTRAR-S est le Topic Model de Griffith, Steyvers et Tenenbaum [11]. Son principal avantage est d'extraire d'un corpus des topics, c'est-à-dire des groupes de mots associés qui relèvent d'un même contexte sémantique et contribuent à l'évocation d'un même thème

ou une même situation. En outre, cette extraction de topics s'accompagne d'une indexation pondérée des documents du corpus à chacun de ses topics, ce qui constitue une première étape du filtrage.

Le Topic Model appartient à un courant de recherche qui applique les statistiques bayésiennes à la découverte de la sémantique des documents et qui a été initié par Hofmann [15]. Ce modèle génératif représente la signification des documents comme une série de topics sélectionnés en fonction de leur distribution probable en mémoire sémantique, et représente la forme de surface du document comme une suite de mots sélectionnés en fonction de leur distribution de probabilité au sein des topics. Appliqué à un corpus de document, le but du modèle est donc de déterminer la meilleure distribution de topics et de mots dans les topics, celle qui est la plus probable étant donné la fréquence de co-occurrence des mots constatés dans le corpus. Le modèle utilise l'algorithme LDA (Latent Dirichlet Allocation) de nature bayésienne non paramétrique [12]. LDA considère la distribution des mots dans un topic donné comme un échantillon d'une distribution de Dirichlet qui est elle-même la « conjugate prior » de la distribution multinomiale représentant la distribution des mots dans un topic. De la même façon, LDA considère la distribution des topics dans un document donné comme une autre distribution de Dirichlet, « conjugate prior » de la distribution multinomiale représentant la distribution des topics dans un document.

À partir d'une distribution multinomiale pour la fonction « likelihood » et d'une distribution de Dirichlet pour la « conjugate prior », on obtient la distribution a posteriori des variables recherchées. Dans le cadre du "Topic Model", les variables observées sont les occurrences des mots dans les documents et les variables recherchées sont les affectations de chacun des mots à un topic. L'échantillonnage de Gibbs, un algorithme de type MCMC (Markov chain Monte Carlo), permet à partir des mots observés de découvrir progressivement les topics apparaissant dans les documents. Les chercheurs ont montré que la distribution des mots dans les topics était une chaîne de Markov ergodique. Après quelques milliers d'itérations, la distribution des mots parmi les topics atteint un état stationnaire.

Le principal avantage du Topic Model est que les topics extraits sont des structures sémantiques directement interprétables. De ce fait, il est aisé de sélectionner une série de topics pour filtrer les documents d'un corpus qui traitent de ces topics et de ceux-là seulement. Des procédures de sélection de topics spécialisés extraits par LDA et de calcul de similarité sur ces sous-espaces ou bases sémantiques pourront alors être évaluées dans le cadre du système FILTRAR-S. En effet, une fois extraits à partir d'un corpus, procédure qui est gourmande en temps de calcul lorsque le corpus est de grande taille, les topics peuvent être assimilés aux dimensions d'un espace vectoriel. Cette possibilité a d'ailleurs été signalée par les auteurs du modèle [13], mais à notre connaissance elle n'a pas encore été véritablement appliquée. Par calcul de similarité elle devrait permettre le filtrage et la récupération de documents qui n'étaient pas présents dans le corpus d'apprentissage, mais dont le contenu est similaire à des documents types. Une procédure interactive de sélection des topics par proposition de mots associés à l'utilisateur est également à l'étude afin de permettre l'enrichissement des requêtes et la sélection des documents pertinents dans un topic. Ces procédures complémentaires devraient permettre au module de filtrage du système FILTRAR-S de répondre aux besoins d'utilisateurs d'une part pour le filtrage sémantique des documents en fonction veille et d'autre part pour la recherche d'informations. Pour l'heure, ces procédures complémentaires ne sont pas encore opérationnelles, mais nous avons déjà franchi une étape importante qui est la réalisation d'un système capable d'extraire des topics et d'indexer les documents du corpus analysés à chacun de ces topics. Les résultats d'une première mise en œuvre du système sont présentés dans le paragraphe qui suit.

3.3 Premiers résultats du module filtrage : extraction de topics et classification des documents par topics

Cette première mise en œuvre a consisté à analyser le corpus du journal « Le Monde » pour l'année 2006. L'objectif était d'identifier les différents topics développés dans un large corpus de documents et de classer ces documents en fonction de leurs topics dominants.

Le corpus est fourni sous forme d'un document XML par article. Après un pré-traitement des articles (suppression des mots vides, lemmatisation), nous avons passé les articles dans une mise en œuvre LDA avec échantillonnage de Gibbs, Gibbs++.

3.3.1 Paramétrage

Pour le lancement de l'outil GibbsLDA++, les paramètres suivants ont été utilisés pour cette première expérimentation :

- Alpha : 0,166667 et Beta : 0,1
- nombre de documents : 43627
- nombre de mots après lémmatisation : 208895
- nombre de topics recherchés : 300
- nombre d'itérations : 3000
- distribution minimum des mots dans le topic : 0,001
- nombre de mots dans le topic : 20
- nombre de documents dans le topic : 50

3.3.2 Résultats

Pour chaque topic, on a obtenu une distribution de probabilité pour 20 mots associés et on a récupéré 50 documents du corpus dont le contenu se rapporte principalement au topic, ainsi que la distribution de probabilité d'appartenance de chacun de ces documents au topic. Un échantillon des résultats obtenus est présenté dans les tableaux 1 et 2. Il s'agit du topic n° 12 dont le thème ou contexte sémantique est « la violence, les jeunes,

la police, les banlieues... », ainsi que les mots du topic l'expriment clairement et plus complètement (cf. tableau 1). En outre, les titres des 20 premiers documents sélectionnés pour ce topic, avec leur probabilité de présence dans le topic (cf. tableau 2) permettent de constater que la plupart d'entre eux sont explicitement pertinents par rapport au topic.

Tableau 1 : Les mots du topic n° 12 et leur probabilité d'appartenance au topic

police	0,0303	policier	0,0290
violence	0,0205	jeune	0,0199
quartier	0,0117	banlieue	0,0098
cité	0,0086	intérieur	0,0070
sécurité	0,0058	délinquance	0,0051
urbain	0,0049	incident	0,0049
bande	0,0044	émeute	0,0043
voiture	0,0042	ordre	0,0041
gendarme	0,0039	S-st-Denis	0,0038
nuit	0,0036	incendie	0,0034

Tableau 2 : Les titres des 20 premiers documents du topic n° 12 et leur probabilité d'appartenance au topic

1 [0,4658] Des bouteilles de gaz avec des clous retrouvées à Aulnay-sous-Bois
2 [0,4090] Un premier rapport de l'IGPN écarte la responsabilité des policiers
3 [0,4042] Des bandes affrontent la police dans le Val-d'Oise
4 [0,3985] Nouveaux incidents entre jeunes et policiers dans les Yvelines
5 [0,3866] Violentes émeutes après la mort de deux adolescents dans une collision avec la police
6 [0,3840] La police trouve des armes à feu après un affrontement à Neuilly-Plaisance
7 [0,3651] Bagarres à répétition entre bandes rivales dans le quartier de la gare du Nord à Paris
8 [0,3639] A Paris, la vendetta entre GDN et Def Mafia déborde dans la rue
9 [0,3527] Les chiffres de la police surestiment le phénomène des bandes
10 [0,352684] Des émeutes urbaines à Cergy sont passées inaperçues
11 [0,3477] Violences urbaines à Saint-Dizier
12 [0,3338] Des gardiens de HLM manifestent leur ras-le-bol après des agressions
13 [0,3310] Les quartiers nord d'Aulnaysous surveillance par peur de nouveaux affrontements
14 [0,330 Emeutes : pourquoi 2007 diffère de 2005
15 [0,3290] Quatorze personnes en garde à vue après des bagarres entre bandes rivales à Paris
16 [0,3283] Provocations policières
17 [0,3197] Près de 730 voitures ont été brûlées au cours de la nuit suivant l'élection
18 [0,3195] Des bandes rivales et la police se sont affrontées à Cergy sans faire événement
19 [0,3167] Treize personnes mises en examen dans l'enquête sur les émeutes de Saint-Dizier
20 [0,31263] L'adolescent renversé par une voiture de police samedi à Marseille est mort

On constate que pour le topic présenté dans le tableau 1, seuls les documents 30, 42 et 49 sont non pertinents par rapport au topic, ce qui équivaut à un taux d'erreurs de 6%. Par ailleurs, une analyse de l'ensemble des 126 topics extraits du corpus nous a permis de constater que 4 d'entre eux rassemblent des mots qui ne permettent pas de dégager un thème cohérent et interprétable. Quatre autres topics rassemblent des mots qui appartiennent visiblement à deux thèmes indépendants l'un de l'autre. Par exemple, le topic n° 14 contient les mots [tabac fumer cigarette fumeur descente ski alpin slalom autrichien skieur …].

On estime d'autre part que 38% des topics extraits expriment des thèmes généraux. Par exemple, le topic n° 108 contient les mots : [université école étudiant élève éducation enseignant enseignement scolaire professeur …]. À l'inverse, 13,5% des topics concernent un fait unique et précis qui a défrayé la chronique. Par exemple, le topic n° 29 concerne exclusivement l'affaire Maddie McCann et contient entre autres les mots [portugais McCann portugal Maddie Madeleine Kate Gerry disparition ...].

Ces variations estimées du degré de généralité des topics pour un même paramétrage des algorithmes révèlent l'influence de la composition du corpus traité et de la redondance du contenu des documents qu'il contient. Ces éléments doivent être pris en compte pour le paramétrage de l'algorithme en fonction des objectifs et devraient faire l'objet de tests pour optimiser le rendement de l'algorithme. Griffith et Steyvers [25] soulignent à ce propos qu'une augmentation de la valeur du paramètre Béta augmente la granularité des topics, tout comme le nombre de topics recherchés. Dans leur application qui concernait la classification par topic de 28 154 articles scientifiques à partir de leurs résumés, Griffith et Steyvers ont testé jusqu'à 1000 topics, mais estiment que le chiffre de 300 topics est optimal.

3.3.3 Conclusions de l'expérimentation

La conclusion de notre première expérimentation du module est que les algorithmes utilisés sont efficaces, mais que différents paramétrages doivent maintenant être testés afin que la granularité des topics extraits soit optimale par rapport aux objectifs de l'utilisateur. En outre, il nous faut maintenant mettre en œuvre les procédures complémentaires de calcul de similarité et d'expansion de requête. Ainsi le module de filtrage pourra être utilisé à la

fois pour filtrer les documents extérieurs au corpus d'apprentissage dans le cadre d'une fonction veille, et pour enrichir de mots clés pertinents les requêtes d'utilisateurs dans le cadre d'une application de type moteur de recherche, pour enrichir des requêtes d'utilisateur.

3.4 La fonction recherche :

Une fonction recherche a été spécialement développée pour le module filtrage et elle est en cours d'amélioration. Actuellement, elle permet de sélectionner les topics les plus fortement associés à un mot, ceux pour lesquels les mots ont une forte probabilité de présence, et renvoie les documents les plus fortement associés aux topics sélectionnés, ceux pour lesquels les topics sélectionnés sont les plus probablement représentatifs.

Les seuils de probabilité pour lesquels les mots et les documents sont considérés comme fortement associés aux topics sont paramétrables en fonction de critères statistiques. Un premier paramètre sélectionne pour chaque topic un seuil de probabilité spécifique pour lequel un mot du corpus analysé est considéré comme présent dans le topic. Ce seuil se fixe sur une base statistique en fonction de la proportion de mots du corpus que l'on veut inclure dans les topics. Un second paramètre effectue la même opération pour la sélection des documents associés aux topics sélectionnés. Par défaut, le système élimine les valeurs de probabilité les plus basses, celles pour lesquelles les différents échantillons de mots et de documents qui sont conservés pour chaque topic ne représentent qu'environ 3 % de ceux qui sont disponibles dans le corpus analysé.

3.5 Premiers essais de la fonction recherche :

- Corpus : « Le Monde 1 mois 11/2007», 3009 documents :
- Paramètres LDA : alpha 0,5 ; beta 0,1 ; Topic 100 ; iteration 100
- Paramètres d'élagage des topics: Paramétrage par défaut conservant pour chaque topic un échantillon d'environ 3 % des mots et des documents du corpus

- Essai 1 : mot « voile »

Deux topics sélectionnés :

a) Un topic dont les mots les plus probables sont : « monde novembre course mille départ… », le mot voile est en position 9 du topic (p= 0,0077)

Les dix premiers documents sélectionnés pour le topic 1 concernent tous les courses de voiliers. La précision à 10 est donc 1 pour ce thème. Ils sont classés par ordre de pertinence en fonction de la valeur de probabilité qui lie chaque document au topic. Trois documents, dont les deux premiers sélectionnés par ordre de pertinence, ne contiennent pas le mot « voile ». Parmi les huit documents suivants qui sont sélectionnés un seul n'est absolument pas pertinent (il concerne les couses cyclistes), et deux autres concernent la marine, mais pas spécialement les course de voilier. Pour les rangs de pertinence suivant, la thématique des documents sélectionnés devient aléatoire et ne concerne apparemment plus les voiliers.

b) Un topic dont les mots les plus probables sont : « musulman monde église catholique chrétien… », le mot voile est en position 115 du topic (p= 0,001)

Les quinze premiers documents sélectionnés par ordre de pertinence pour ce topic concernent la religion, et les suivants concernent d'autres thématiques. A deux exceptions près, aucun de ces quinze documents sur la religion ne contient le mot voile. Les deux documents qui contiennent le mot voile concerne le voile islamique. L'utilisateur intéressé spécialement par le problème du voile islamique devra donc limiter sa recherche aux seuls documents du topic qui contiennent le mot « voile ».

- Essai 2 : mot « vol »

Deux topics sélectionnés :

a) « Britannique compagnie avion Londres air transports … », le mot vol est en position 30 du topic (p= 0,0036)

Les dix premiers documents sélectionnés par ordre de pertinence pour ce topic sauf 1 concernent le transport aérien, l'exception concerne le transport ferroviaire. La précision à 10 pour une requête concernant ce thème est donc de 0,9. Parmi ces dix premiers documents sélectionnés, cinq ne contiennent pas le mot « vol » tout en étant pertinent par rapport au thème du transport aérien. Parmi les sept documents sélectionnés suivants , deux ne sont pas pertinents et les documents suivants ne concernent apparemment plus le transport aérien.

b) « arme trafic mafia drogue contre antiquaire… », le mot vol est en position 23 du topic (p= 0,0034)

Parmi les documents sélectionnés par ordre de pertinence sous ce topic, les trois premiers concernent le vol d'oeuvre d'art, et, à l'exception d'un document sur les vol de métaux, les sept suivants concernent la mafia, le trafic de drogue et la délinquance en banlieue, c'est-à-dire des problématiques moins directement liée à l'acte de vol proprement dit mais plutôt au banditisme. La même pluralité de thématiques liées au banditisme se retrouve dans les sept documents suivants. Ensuite, les documents ont des thématiques différentes. Il est à noter ici que les seuls documents liés spécifiquement au vol contiennent explicitement ce mot et sont sélectionné dans le topic.

- **Conclusions :**

La pertinence des premiers documents associés aux topics sélectionnés par la fonction recherche est relativement bonne, même si davantage de précision est

parfois obtenu en exigeant aussi que le mot recherché soit aussi présent dans les textes.
Le problème reste celui de l'évaluation du taux de rappel, c'est-à-dire la prise en compte du nombre total de documents pertinents qui figurent dans le corpus et dont une partie seulement est sélectionnée dans le topic.

3.6 Développements futurs :

1.1.1 Sélection interactive :

Actuellement beaucoup de documents non pertinents restent sélectionnés par les topics, quoique majoritairement à la suite des documents pertinents, et certains topics peuvent clairement se subdiviser en plusieurs sous-topics. Une solution simple consisterait à sélectionner pour chaque topic les mots du topics qui doivent apparaître dans les documents renvoyés à l'utilisateur.

La procédure d'utilisation de la fonction recherche devient alors : 1) entrer un (ou plusieurs) mots dans le champ "search", 2) sélection d'un ou plusieurs topics pertinents parmi ceux renvoyés pour le mot entré, 3) sélection par l'utilisateur de quelques mots pertinents dans chaque topic sélectionné en cliquant sur certains mots parmi les premiers de chaque topic (par exemple les 20 premiers mots). Puis deux possibilités:

a) Renvoyer à l'utilisateur les documents du topic qui contiennent les mots sélectionnés

b) Renvoyer à l'utilisateur les documents du topic qui contiennent seulement quelques-uns des mots sélectionnés : un, deux, trois ... mots, selon le paramètre fixé par l'utilisateur.

1.1.2 HLDA :

Ces premiers résultats montrent que LDA est un algorithme efficace qui extrait des topics conformes à nos intuitions, et dans lesquels les mots regroupés ont un rapport de sens avec des situations connues. Cependant, les premiers résultats suggèrent également que des regroupements de plusieurs topics appartenant à une même situation sont possibles et permettraient de simplifier les résultats. C'est pourquoi nous avons décidé de mettre en oeuvre l'algorithme Hierachical LDA (hLDA) qui vise précisément à opérer ces regroupements entre topics et à construire une hiérarchie de topics. Un autre avantage important de l'algorithme est que le nombre de topics n'est plus à fixer au départ. En effet, LDA contraint l'utilisateur à fixer a priori et à l'aveugle le nombre de topics qui seront extraits du corpus.

4. Extraction d'information non supervisée

4.1 Problématique

L'extraction d'information, longtemps envisagée sous le seul angle du paradigme des conférences MUC (Message Understanding Conferences) 5, prend depuis quelques temps des formes plus diverses. À côté de la tâche consistant à extraire de textes les éléments d'information venant occuper un rôle bien défini dans une structure informationnelle donnée *a priori* (souvent appelée *template*), sont apparues des tâches moins contraintes, notamment du point de vue de la spécification des informations à extraire. Traditionnellement, celles-ci sont décrites sous la forme d'une configuration de relations entre entités, chaque relation étant spécifiée par un modèle élaboré manuellement (généralement sous la forme d'un ensemble de règles) ou par un ensemble d'exemples de relations en contexte permettant d'apprendre un modèle, souvent de nature statistique. Une première extension de ce mode de spécification se caractérise par la donnée d'exemples ou de modèles sous-déterminés, c'est-à-dire nécessitant d'être étendus pour être exploités à des fins d'extraction d'information. Cette problématique se retrouve dans les approches fondées sur l'amorçage 5 ou plus récemment, dans celles relevant de la notion de supervision distante (« distant supervision ») 5, approche particulièrement mise à l'honneur par la tâche Knowledge Based Population (KPB) de l'évaluation Text Analysis Conference (TAC) 5 et dans laquelle les exemples de relations se limitent à des couples d'entités, sans instanciation en corpus.

Un pas supplémentaire dans le relâchement de la supervision est accompli avec la notion d'extraction d'information non supervisée que nous avons adoptée pour FILTRAR-S. Dans ce cas, la supervision se limite en effet à fixer des contraintes sur les relations extraites, portant typiquement sur le type des entités liées, sans qu'aucun exemple ou modèle de ces relations ne soit fourni. Cette approche, incarnée par 5 ou 5, est particulièrement adaptée à un contexte de veille, contexte dans lequel les entités dignes d'intérêt et les sources d'information sont connues tandis que les relations entre ces entités sont les informations recherchées sans avoir d'*a priori* très précis les concernant.

4.2 Le module d'extraction d'information non supervisée de FILTRAR-S

Nous rappelons ici les grandes lignes du module d'extraction d'information non supervisée de FILTRAR-S tel qu'il a été décrit dans 5 et dans 5. Ce module assume deux fonctions principales. Il commence par extraire des textes des relations entre entités nommées pour ensuite les regrouper selon deux dimensions : regroupement d'abord sur le plan thématique, puis sur le plan sémantique. Le premier regroupement vise à rassembler les relations intervenant dans un même thème tandis que le second a

pour objectif de faire émerger au sein d'un thème les relations équivalentes sur le plan sémantique, c'est-à-dire les relations en état de paraphrase.

Les relations extraites des textes sont caractérisées par trois grandes catégories d'information permettant tout à la fois de les définir et de fournir les éléments nécessaires à leurs regroupements :

- un couple d'entités nommées (E1 et E2). Dans les expérimentations menées, nous nous sommes restreints aux entités de type PERSONNE (PERS), ORGANISATION (ORG) et LIEU ;
- une caractérisation linguistique de la relation. Il s'agit de la façon dont la relation est exprimée linguistiquement. Chaque relation étant extraite sur la base de la présence dans une phrase d'un couple d'entités nommées correspondant aux types ci-dessus, sa caractérisation linguistique comporte trois parties :
  - *Cpre* : la partie de la phrase précédant la première entité (E1) ;
  - *Cmid* : la partie de la phrase se situant entre les deux entités ;
  - *Cpost* : la partie de phrase suivant la seconde entité (E2) ;
- un contexte thématique. Celui-ci a pour rôle de rendre compte du thème associé à la relation. Il prend la forme des mots pleins du segment thématique d'où la relation considérée est extraite.

On pourra se reporter à la figure 2 pour un exemple de relation extraite et de ses constituants.

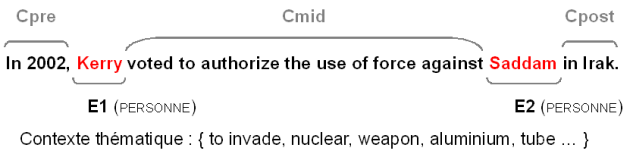

FIG. 2 : exemple de relation extraite[1]

La phase d'extraction des relations s'appuie sur un prétraitement linguistique des textes permettant de mettre en évidence dans les textes les informations nécessaires à la définition des relations. Ce prétraitement comporte donc une reconnaissance des entités nommées pour les types d'entités visés, une désambiguïsation morpho-syntaxique des mots ainsi que leur normalisation, et enfin au niveau textuel, une segmentation thématique. Cette dernière est réalisée par l'outil LCseg 5 tandis que les autres traitements s'appuient sur les outils d'OpenNLP 5.

Les regroupements thématique et sémantique reposent quant à eux sur le même algorithme de clustering, le *Markov Clustering* 5, mais portent sur des dimensions différentes des relations. Le regroupement thématique s'effectue ainsi à partir d'un graphe de similarité calculé en appliquant la mesure *cosinus* aux contextes thématiques des relations alors que le regroupement sémantique, au sein de chaque cluster thématique formé, est réalisé à partir d'un graphe de similarité calculé sur la base des caractérisations linguistiques des relations. Dans ce dernier cas, les expérimentations menées ont montré que la mesure *cosinus* prenant en compte les trois composantes de cette caractérisation donne les meilleurs résultats. Dans le cas du regroupement thématique, le nombre important de similarités à calculer[2] est contrebalancé par le recours à l'algorithme *All Pairs Similarity Search* (APSS) 5 qui permet, moyennant la fixation d'un seuil de similarité minimale, de calculer efficacement la mesure *cosinus* entre les contextes thématiques considérés comme des sacs de mots.

4.3 Filtrage des relations

Dans le module d'extraction d'information non supervisée tel qu'il est décrit à la section précédente, les contraintes pesant sur les relations sont très limitées. Sont ainsi extraites les relations correspondant à tout couple d'entités nommées dont les types correspondent aux types ciblés, avec pour seules restrictions la cooccurrence de ces entités dans une même phrase et la présence d'au moins un verbe entre les deux. Le tableau 3 donne le volume des relations ainsi extraites de la sous-partie du corpus AQUAINT-2 constituée de 18 mois du journal *New York Times* sur laquelle se sont appuyées les expérimentations que nous avons menées.

TAB. 3 : volumétrie des relations extraites

PERS – PERS	175 802
PERS – ORG	126 281
PERS – LIEU	152 514
ORG – ORG	77 025
ORG – LIEU	71 858
ORG – PERS	73 895
LIEU– ORG	57 092
LIEU– PERS	78 845
LIEU– LIEU	116 092

Un examen de ces relations montre cependant qu'un nombre très significatif des relations ainsi extraites ne correspondent pas à de véritables relations entre les entités impliquées. Il semble donc que cette stratégie basique d'extraction, qui peut donner des résultats intéressants dans des domaines de spécialité[3], ne soit pas suffisamment sélective en domaine ouvert. Nous avons donc cherché à la compléter par un processus de filtrage spécifique visant, comme dans 5, à déterminer si deux entités dans une

[1] L'exemple est donné en anglais car nos expérimentations ont été réalisées dans cette langue.

[2] Il est en théorie nécessaire d'évaluer la similarité de tous les contextes thématiques deux à deux, ce qui peut poser problème lorsque l'on souhaite traiter plusieurs centaines de milliers de relations.

[3] Le travail rapporté dans 5 montre que dans le domaine médical, les relations extraites sur la base de cette stratégie sont correctes dans 79% des cas.

phrase sont ou ne sont pas liées par une relation, sans *a priori* sur la nature de cette relation.

4.3.1 Filtrage heuristique

Dans une perspective exploratoire, nous avons défini un nombre restreint d'heuristiques de filtrage et analysé leur impact. Ces heuristiques sont au nombre de trois :

- la suppression des relations comportant entre leurs deux entités un verbe exprimant un discours rapporté (dans le cas présent, la liste se limite aux verbes *to say* et *to present*). Ceci vise à éviter d'extraire une relation entre les entités *Homgren* et *Allen* dans l'exemple suivant:

  Holmgren said Allen was more involved with the team …
- un nombre de mots entre les deux entités limité à 10. Au-delà de cette limite empirique, le nombre des relations effectives entre les deux entités devient en effet très faible ;
- la limitation à 1 du nombre de verbes entre les deux entités, sauf si ces verbes ont valeur d'auxiliaire (*be*, *have* et *do*).

L'application de ces heuristiques aux relations extraites a globalement pour conséquence de réduire leur volume d'environ 50%. Le tableau 4 illustre plus précisément ce ratio pour chaque type de relations considéré à partir d'un échantillon de 8 000 relations pour chaque type.

TAB. 4 : effet de l'application des heuristiques de filtrage

Type de relations	filtrées/ gardées	discours rapporté	distance	1 seul verbe
LOC-LOC	4287/3713 (46%)	440	3548	2763
LOC-ORG	4097/3903 (49%)	488	3224	2650
LOC-PER	4790/3210 (40%)	1636	3352	2638
ORG-LOC	4225/3775 (47%)	643	3324	2869
ORG-ORG	4169/3831 (48%)	627	3123	2810
ORG-PER	4541/3459 (43%)	1541	3155	2859
PER-LOC	4209/3791 (47%)	905	3199	2813
PER-ORG	3888/4112 (51%)	952	2742	2566
PER-PER	4444/3556 (44%)	1290	3109	2741

Chacune des trois dernières colonnes donne le nombre de relations filtrées par l'heuristique considérée, sachant qu'une relation peut-être filtrée par plusieurs heuristiques. La deuxième colonne fournit quant à elle le nombre de relations filtrées et le nombre de celles qui sont conservées, avec le pourcentage que représentent ces dernières. L'heuristique la plus filtrante est clairement celle de la distance entre entités mais celle limitant le nombre de verbes a également un impact très significatif.

Néanmoins, ces ratios de filtrage doivent être mis en parallèle avec une évaluation de l'efficacité des heuristiques correspondantes en termes de sélection des relations correctes. Pour ce faire, nous avons choisi au hasard 50 relations de chaque type et nous avons procédé à une annotation manuelle de leur validité au moyen de l'interface de la figure 3, interface Web générée à partir de la représentation XML des relations par des règles XSLT.

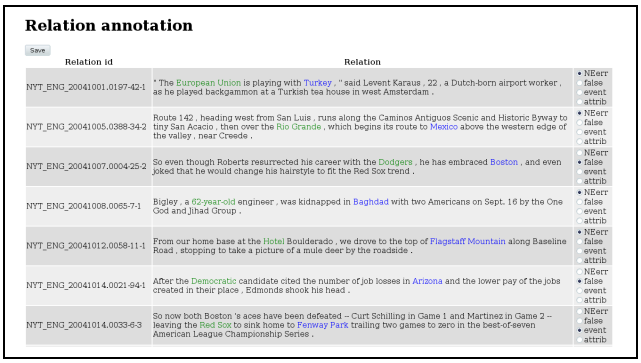

FIG. 3 : interface d'annotation des relations

Le tableau 5 donne le résultat de cette évaluation montrant que globalement, le taux de fausses relations parmi les relations filtrées est assez élevé pour tous les types de relations mais que parmi les relations conservées, certains types de relations, en particulier toutes les relations ayant un lieu comme première entité nommée, se caractérisent par un taux de fausses relations encore très important.

TAB. 5 : évaluation du filtrage par les heuristiques

Type de relations	Filtrées		Gardées	
	correctes	fausses	correctes	fausses
LIEU-LIEU	1	49 (98%)	9 (18%)	41
LIEU-ORG	4	46 (92%)	8 (16%)	42
LIEU-PER	3	47 (94%)	2 (4%)	48
ORG-LIEU	7	43 (86%)	14 (28%)	36
ORG-ORG	6	44 (88%)	20 (40%)	30
ORG-PER	4	46 (92%)	20 (40%)	30
PER-LIEU	13	37 (74%)	40 (80%)	10
PER-ORG	12	38 (76%)	40 (80%)	10
PER-PER	5	45 (90%)	14 (28%)	36

Ce constat n'est d'ailleurs pas surprenant dans la mesure où la première entité d'une relation occupe souvent un rôle d'agent alors que les lieux apparaissent le plus fréquemment comme des circonstants. Compte tenu de cette observation, nous avons choisi d'écarter systématiquement les relations ayant un lieu comme première entité dans la suite des traitements.

4.3.2 Filtrage par apprentissage

L'évaluation précédente a mis en évidence l'intérêt des heuristiques testées pour écarter les mauvaises relations mais a également montré leur insuffisance pour conserver une proportion significative des relations correctes. Nous avons donc choisi d'adjoindre à ces heuristiques un

module de filtrage reposant sur un classifieur statistique décidant si une relation extraite est véritablement sous-tendue par une relation effective entre ses entités.

La première tâche pour ce faire a été de construire un corpus de référence en annotant manuellement un ensemble de relations au moyen de l'interface de la figure 3. Plus précisément, 200 relations ont été sélectionnées au hasard et annotées pour chacun des 6 types de relations finalement considérés. L'annotation distinguait les relations correctes, les relations incorrectes du fait d'un problème de reconnaissance des entités nommées et les relations fausses du fait de l'absence de relation effective.

TAB. 6 : résultat de l'annotation manuelle des relations

Type de relations	correctes	erreurs EN	fausses
ORG-LOC	38% (77)	18% (35)	44% (88)
ORG-ORG	39% (78)	14% (28)	47% (94)
ORG-PER	36% (72)	18% (36)	46% (92)
PER-LOC	51% (102)	31% (62)	18% (36)
PER-ORG	60% (120)	18% (36)	22% (44)
PER-PER	41% (82)	20% (39)	40% (79)
Tous	44% (531)	20% (236)	36% (433)

Les relations incorrectes du fait des entités nommées représentent environ 20% de l'ensemble et ont été laissées de côté pour l'entraînement et le test des classifieurs. Le corpus résultant se compose donc de 964 relations, 531 étant correctes et 433 étant fausses, ce qui constitue un ensemble suffisamment équilibré pour ne pas poser de problème spécifique pour l'apprentissage des modèles statistiques.

Plusieurs de ces modèles ont été testés en se concentrant d'abord sur des modèles exploitant un ensemble de caractéristiques locales non structurées. Classiquement, nous avons ainsi entraîné un classifieur bayésien naïf, un classifieur de type maximum d'entropie (MaxEnt), un arbre de décision et un classifieur fondé sur les Machines à Vecteurs de Support (SVM). Pour les trois premiers, nous nous sommes appuyés sur l'implémentation fournie par la boîte à outils MALLET 5 tandis que pour le dernier, nous avons eu recours à l'outil SVM^{light} 5. Ces différents classifieurs ont été entraînés en utilisant le même ensemble de caractéristiques. Ces caractéristiques reprennent celles utilisées classiquement pour l'extraction de relations, à l'instar de 5 :

- le type des entités nommées E1 et E1 ;
- la catégorie morpho-syntaxique des mots situés entre les deux entités, avec une caractéristique binaire pour chaque couple (*position dans la séquence*, *catégorie*), ainsi que les bigrammes de catégories morpho-syntaxiques entre E1 et E2, avec une caractéristique binaire pour chaque triplet (*position i*, cat_i, cat_{i+1}) ;
- la catégorie morpho-syntaxique des deux mots précédant E1 et des deux mots suivant E2, à la fois en tant qu'unigrammes et en tant que bigrammes ;
- la séquence des catégories morpho-syntaxiques entre E1 et E2. Chaque séquence possible de 10 catégories est encodée comme une caractéristique binaire ;
- le nombre de mots entre E1 et E2 ;
- le nombre de signes de ponctuation (virgule, guillemet, parenthèse …) entre E1 et E2.

Compte tenu de la taille relativement réduite du corpus pour chaque type de relations, nous avons choisi d'évaluer ces différents classifieurs en faisant appel à la technique classique de la validation croisée. Le corpus annoté a ainsi été découpé en 10 parties égales, 9 parties étant utilisées pour l'entraînement des classifieurs, la partie restante pour le test, le processus étant mené 10 fois afin que chaque partie serve à la fois pour l'entraînement et le test. Les résultats donnés par le tableau 7 sont des moyennes sur ces 10 itérations pour les mesures suivantes :

$$Accuracy = \frac{nombre\ de\ relations\ bien\ classées}{nombre\ total\ de\ re}$$

$$Précision = \frac{nombre\ de\ relations\ correctes\ ide}{nombre\ de\ relations\ identifie}$$

$$Rappel = \frac{nombre\ de\ relations\ correctes\ identi}{nombre\ de\ relations\ c}$$

(1)

la F1-mesure étant la moyenne harmonique de la précision et du rappel.

TAB. 7 : évaluation des classifieurs statistiques

Modèle	Accuracy	Précision	Rappel	F1-mesure
Bayésien naïf	0,637	0,660	0,705	0,682
MaxEnt	0,650	0,665	0,735	0,698
Arbre de décision	0,639	0,640	0,784	0,705
SVM	0,732	0,740	0,798	0,767
5	/	0,883	0,452	0,598

Ce tableau montre en premier lieu que les meilleurs résultats sont obtenus par le classifieur de type SVM, ce qui n'est pas surprenant au vu des travaux réalisés de façon générale sur l'extraction de relations. On notera également un certain équilibre entre la précision et le rappel et ce, pour tous les types de classifieurs. Enfin, ces résultats se comparent favorablement à ceux de 5 sur le même sujet comme le montre la dernière ligne du tableau 7. Dans ce dernier cas, le profil des résultats est un peu différent puisque la précision est plus forte que la nôtre mais le rappel très largement inférieur. Il faut néanmoins préciser que dans 5, les relations extraites peuvent faire intervenir des entités plus générales que des entités nommées, ce qui est *a priori* un facteur de difficulté.

4.3.3 Modèle de séquence pour le filtrage par apprentissage

À l'instar de 5, nous avons également testé un classifieur prenant en compte la notion de séquence en nous appuyant sur les *Champs Conditionnels Aléatoires* (CRF). Dans ce cas, la tâche considérée n'est plus directement une tâche de classification des relations mais prend la forme d'un étiquetage, illustré par la figure 4.

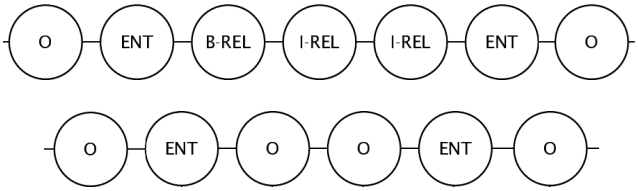

FIG. 4 : étiquetage des relations par un modèle CRF

Plus précisément, il s'agit d'étiqueter chaque mot d'une phrase par l'une des quatre étiquettes suivantes, suivant en cela le modèle IOB introduit par 5 :

- O : mot de la phrase en dehors d'une relation ;
- ENT : élément d'une entité nommée définissant une relation potentielle (E1 ou E2) ;
- B-REL : premier mot d'une relation suivant E1 ;
- I-REL : mot faisant partie d'une relation.

Dans ce schéma, une relation est jugée correcte lorsque l'étiquetage suit la première configuration de la figure 4 (avec un nombre de I-REL variable selon la relation) tandis qu'elle est jugée fausse lorsque l'étiquetage produit la seconde configuration.

Comme les classifieurs de la section précédente, ce modèle à base de CRF s'appuie sur un ensemble de caractéristiques :

- la catégorie morpho-syntaxique du mot courant, du mot précédent et du mot suivant ;
- les bigrammes de catégories morpho-syntaxiques $<cat_{i-1}, cat_i>$, avec i=-1,0,1 (0 : mot courant ; -1 : mot précédent ; 1 : mot suivant) ;
- le type d'entité nommée du mot courant et de chacun des 6 mots le précédant et le suivant. Ce type peut avoir une valeur NIL lorsque le mot ne fait pas partie d'une entité nommée.

TAB. 8 : évaluation du modèle CRF

Modèle	Accuracy	Précision	Rappel	F1-mesure
SVM	0,732	0,740	0,798	0,767
CRF	0,745	0,762	0,782	0,771

Le tableau 8 montre les résultats obtenus par ce modèle CRF, implémenté au moyen de l'outil Wapiti 5, suivant les mêmes modalités de validation croisée utilisées pour les classifieurs de la section précédente. La comparaison avec le meilleur de ceux-ci met en avant une légère supériorité du modèle à base de CRF, avec toujours le même équilibre entre précision et rappel. C'est donc ce modèle que nous avons retenu pour le filtrage des relations dans le cadre de notre processus d'extraction d'information non supervisée.

4.3.4 Application du filtrage des relations

L'extraction des relations telle que nous l'avons envisagée précédemment se compose des 3 étapes suivantes, appliquées successivement :

- une extraction initiale ne posant comme contraintes que la cooccurrence dans une phrase d'entités nommées relevant d'un ensemble donné de types et la présence d'au moins un verbe entre les deux ;
- l'application de 3 heuristiques simples permettant d'écarter avec une bonne précision un grand nombre de relations fausses ;
- l'application d'un modèle de filtrage à base de CRF permettant de discriminer plus finement les relations correctes.

Le constat de la présence dans nos relations filtrées d'un certain nombre de relations identiques, pour une part issues d'articles sur un même sujet ou d'articles correspondant à des rubriques très formatées, nous a conduit à compléter ce processus de filtrage par un dédoublonnage final visant à éliminer ces relations redondantes. Pour ce faire, nous détournons le processus de regroupement sémantique en fixant une valeur de similarité de 1.0 à la mesure *cosinus* appliquée à une représentation « sac de mots » des phrases. Une relation est ensuite choisie pour représenter chaque cluster de relations identiques ainsi formé.

Le tableau 9 illustre l'application de la totalité des étapes de filtrage aux relations du tableau 3. On constate que ce filtrage laisse de côté un grand nombre des relations extraites initialement mais que le volume des relations restantes est *a priori* suffisant pour alimenter efficacement les étapes suivantes de notre processus d'extraction d'information non supervisée. Par ailleurs, comme 5, nous nous situons dans un contexte de traitement de volumes textuels importants caractérisés par une certaine redondance informationnelle conduisant à privilégier la précision des processus d'extraction afin d'éviter un bruit trop important.

5. Conclusion et perspectives

Nous avons présenté ici les derniers développements de FILTRAR-S concernant le filtrage et la fouille des textes orientée par les relations entre entités. Dans le cas du filtrage, une validation de l'utilisation des topics pour le filtrage de documents est en cours, avec en particulier une évaluation se fondant sur les données de campagnes d'évaluation de référence de type CLEF. Du côté de la fouille de textes, outre l'approfondissement des résultats déjà obtenus, un travail d'intégration plus étroite est encore à mener, en particulier afin que le module de question-réponse soit à même d'exploiter les annotations produites par le module d'extraction d'information non supervisée.

TAB. 9 : volume des relations à l'issue de chacune des étapes de filtrage

Type des relations	ORG-LOC	ORG-ORG	ORG-PERS	PERS-LOC	PERS-ORG	PERS-PERS
extraction initiale	71 858	77 025	73 895	152 514	126 281	175 802
heuristiques	33 505	37 061	32 033	72 221	66 035	78 530
classifieur CRF	16 700	17 025	12 098	55 174	50 487	42 463
dédoublonnage	15 226	13 704	10 054	47 700	40 238	38 786

	(21%)	(18%)	(14%)	(31%)	(32%)	(22%)

Références

[1] Campion, N., Martins, D., Whilhem, A. (2009). Contradictions and Predictions: Two Sources of Uncertainty that Raise the Cognitive Interest of Readers, *Discourse Processes*, 46:1–28.

[2] Hare, M., Jones, M., Thomson, C. , Kelly, S. , McRae, K. (2009). Activating event knowledge. *Cognition*, 111 (2), 151-167.

[3] Campion N., Rossi J.-P., Le Ny, J.-F., Declercq, C. (2006). Action schema, a basic knowledge structure accessed to provide meaning relations to words. *Sixteenth Annual Meeting of the Society for Text and Discourse*, 13-15 juillet, Minneapolis, USA.

[4] Campion, N. & Rossi, J.-P. (2008). Les schémas d'actions : structure et fonction d'un composant de la mémoire sémantique. *Communication orale au colloque de L'Arco*, Lyon.

[5] Campion, N., Closson, J., Carcenac, J., Ferret O., Grau, B., & Shin, J. (2009). Modélisation de la mémoire sémantique et compréhension des messages par Filtrar-S : Un cyber-outil pour la sécurité globale. *Actes du troisième Workshop Interdisciplinaire sur la Sécurité Globale (WISG 2009)*, Troyes, France.

[6] Campion, N., & Rigalleau, F. (2008) Les atteintes fonctionnelles précoces de la mémoire sémantique dans la maladie d'Alzheimer : Diagnostic et remédiation cognitive, *colloque ARCo'08 - Connaissances : Genèse, Nature et Fonction*, 3-5 décembre, Lyon

[7] Salton G. & McGill M. (1983) *Introduction to Modern Information Retrieval*, McGraw-Hill.

[8] Landauer, T. K. and Dumais, S. T. (1997) A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), 211-240.

[9] Burgess, C. et Lund, K. (2000). The dynamics of meaning in memory. In Dietrich, E., & Markman, A. B. (Eds.), *Cognitive Dynamics: Conceptual Change in Humans and Machines*. Lawrence Erlbaum Associates, Publishers. pp. 117-156.

[10] Jones, M. N., & Mewhort, D. J. K. (2007). Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, 114-1, 1–37.

[11] Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114, 211-244.

[12] Blei, D.M., NG, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

[13] Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis.* Hillsdale, NJ: Erlbaum. pp. 427-448.

[14] Dumais, S. T. (2007). LSA and information retrieval: Getting back to the basics. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis.* Hillsdale, NJ: Erlbaum. pp. 293-321.

[15] Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceeding of the 22nd Annual International SOGOR Conference on Research and Development in Information Retrieval*, 50-57.

[16] Blei, D. M., Griffiths, T. L., & Jordan, M. I. (in press). The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM.*

[17] KINTSCH W., *On the notions of theme and topic in psychological process models of text comprehension*, In W. van Peer and M.M. Louwerse (Eds.), Thematics in psychology and literary studies, p 157–170), Amsterdam: Benjamins, 2002

[18] CAMPION N. and ROSSI J.-P., *Associative and causal constraints in the process of generating predictive inferences*, Discourse Processes, 31(3), p 263–291, 2001

[19] R. Grisham et B. Sundheim. *Design of the MUC-6 evaluation*. 6th conference on Message understanding, 1995.

[20] E. Agichtein et L. Gravano. Snowball: *Extracting relations from large plain-text collections*. 5th ACM International Conference on Digital Libraries, 2000.

[21] M. Mintz, S. Bills, R. Snow et D. Jurafsky. *Distant supervision for relation extraction without labeled data*. ACL-IJCNLP'09, p. 1003-1011, 2009.

[22] H. Simpson, S. Strassel, R. Parker et P. McNamee. *Wikipedia and the Web of Confusable Entities: Experience from Entity Linking Query Creation for TAC 2009 Knowledge Base Population*. 7th Conference on Language Resources and Evaluation (LREC 2010), Malta, 2010.

[23] Y. Shinyama et S. Sekine. Preemptive *Information Extraction using Unrestricted Relation Discovery*. Human Language Technology Conference of the NAACL, 2006, p. 304-311.

[24] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead et O. Etzioni. *Open Information Extraction from the Web*. Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), 2007.

[25] N. Campion, J. Closson, O. Ferret, J. Shin, B. Grau, J.M. Lazard, J.M. Floret, A.J. Mezaour; D. Lahbib et R. Besançon. *Filtrar-S : premiers résultats*. 4ème Workshop Interdisciplinaire sur la Sécurité Globale (WISG 2010), Troyes, France.

[26] N. Campion, J. Closson, O. Ferret, J. Shin, B. Grau, J.M. Lazard, D. Lahbib, R. Besançon, J.M. Floret, A.J. Mezaour et Xavier Tannier. *FILTRAR-S : un outil de filtrage sémantique et de fouille de textes*

pour la veille. 6ème Colloque Veille Stratégique Scientifique & Technologique (VSST'2010), 2010, Toulouse, France.

[27] M. Galley, K. McKeown, E. Fosler-Lussier et H. Jing. *Discourse Segmentation of Multi-Party Conversation*. 41st Annual Meeting on Association for Computational Linguistics, 2003, p. 562-569.

[28] OpenNLP : http://opennlp.sourceforge.net/

[29] S. Van Dongen. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, 2000.

[30] R. J. Bayardo, Y. Ma et R. Srikant. *Scaling Up All-Pairs Similarity Search*. 16th International Conference on World Wide Web (WWW 2007), 2007.

[31] M. Embarek et O. Ferret. *Learning patterns for building resources about semantic relations in the medical domain*. 6th Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.

[32] M. Banko et O. Etzioni. *The Tradeoffs Between Open and Traditional Relation Extraction*. ACL-08, Columbus, Ohio, 2008, p. 28-36.

[33] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[34] T. Joachims. *Making large-Scale SVM Learning Practical*. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[35] L. Ramshaw et M. Marcus. *Text Chunking Using Transformation-Based Learning*. Third ACL Workshop on Very Large Corpora, Cambridge, USA.

[36] T. Lavergne, O. Cappé et F. Yvon. *Practical very large scale CRFs*. 48th Annual Meeting of the Association for Computational Linguistic (ACL 2010), 2010.