GARDER SES INFORMATIONS À JOUR : UN ENJEU STRATÉGIQUE

Marion LAIGNELET
marion.laignelet@irit.fr
IRIT - Équipe IC3
118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9

Résumé

Nous présentons dans cet article un travail concernant les encyclopédies et la problématique de mise à jour de l'information contenue dans ce type de documents généralement volumineux. Ce projet réunit des professionnels de l'édition qui ont exprimé un besoin concret concernant la question de la validité des informations dans les textes et des linguiste-informaticiens à même de proposer des solutions de TAL concrètes pour le développement d'un système d'aide à la mise à jour.

La question de la mise à jour des textes est envisagée à travers la notion d'obsolescence : la recherche des segments textuels contenant une information susceptible d'évoluer dans le temps est alors visé. Cette annotation est basée sur le repérage de traits textuels hétérogènes sémantiquement et à granularité variable. Un système d'apprentissage automatique permet de mettre en lumière les traits et combinaisons de traits linguistiques pertinents dans les segments contenant de l'information susceptible d'être obsolète. Nous montrons finalement que pour une tâche telle que celle de mise à jour, il est important de prendre en considération la thématique des textes.

1 Introduction

Parce que de plus en plus de documents sont mis à disposition sur internet, la question de la validité des informations transmises se pose de manière récurrente. Sur internet, les sources d'information sont nombreuses et les types de documents le sont tout autant. Nous présentons dans cet article un travail concernant les encyclopédies et la problématique de mise à jour de l'information contenue dans ce type de documents.

Le monde de l'édition et plus spécifiquement celui de l'édition encyclopédique connaît une réelle mutation technique. Le nombre de mises en ligne d'encyclopédies généralistes sur Internet en est révélateur. Ainsi, *Wikipédia* est lancé en 2001[1] sur Internet, *Larousse* en 2008[2] ; *Hachette*[3], *Encyclopedia Universalis*[4],

[1] http://fr.wikipedia.org/wiki/Accueil
[2] http://www.larousse.fr/encyclopedie/
[3] http://www.ehmelhm.hachette-multimedia.fr
[4] http://www.universalis.fr/

Encarta[5], ou encore *Encyclopédie gratuite*[6] sont également disponibles sur Internet.

Face à cette réalité, les acteurs de l'édition doivent être le plus réactif possible aux évolutions des connaissances dans tous les domaines qu'ils couvrent et pas seulement dans les sujets d'actualité. Il semble toutefois difficile d'envisager une mise à jour manuelle systématique notamment à cause de la taille de ce type de documents et de la quantité d'information disponible. Et à l'heure où internet permet un accès permanent à l'information, il est d'autant plus important pour les éditeurs de proposer des versions mises à jour en temps réel. Ceci est valable tant pour les versions en ligne que pour les versions papier même si pour ces dernières le problème de la mise à jour est différent, d'un point de vue temporel du moins.

Pour répondre à ce besoin, nous proposons un outil d'aide à la mise à jour de contenus encyclopédiques. Nous partons de l'idée qu'un repérage et/ou une extraction automatique des zones contenant de l'information potiellement obsolète est à même d'aider les rédacteurs dans leur tâche de mise à jour des contenus. Le prototype proposé est basé sur la prise en compte de traits textuels hétérogènes et à granularité variable. Les résultats actuels montrent notamment la nécessité d'adapter les traitements en fonction des rubriques thématiques (histoire, géographie, sciences, médecine, etc.).

Dans la section suivante, nous présentons les moyens actuels mis en oeuvre pour mettre à jour les informations dans les encyclopédies. Dans la section 3, nous présentons des travaux de recherche proches du nôtre comme la recherche de l'évolution dans les textes ou encore la recherche de la nouveauté. Nous focalisons ensuite notre attention sur la notion d'obsolescence de l'information. La section 4 présente l'outil (à l'heure actuelle plutôt un prototype) développé afin d'aider à la mise à jour des informations dans les textes sur la base du repérage automatique de segments potentiellement obsolescents. Enfin, la dernière section insiste sur le fait que pour améliorer les résultats, il faudrait mieux considérer les aspects thématiques des entrées encyclopédiques.

2 Les solutions actuelles de mise à jour de l'information dans les encyclopédies

Cet article s'inscrit dans le cadre d'un projet de recherche qui a pris naissance au sein de l'entreprise d'édition-packaging[7], Initiales (Montpellier), et qui s'est poursuivi avec un partenariat avec les éditions Larousse (Paris).

La question de la mise à jour des données textuelles s'est posée lorsque la société Initiales s'est trouvée en charge d'éditer et de mettre à jour une collection de fiches encyclopédiques éditées par les Éditions ATLAS. Cette collection, diffusée dans le cadre d'*éditions au long cours*[8], est constituée de fascicules thématiques qui sont envoyés par voie postale aux clients sous forme d'abonnement. Ce type de fonctionnement est généralement mis en place pour durer trois ans, voire plus : il ne s'agit pas d'éditer une fois un ouvrage mais d'étaler les éditions et publications des fiches sur un temps déterminé. Étant donné que les abonnements ne débutent pas tous au même moment, une même fiche est susceptible d'être envoyée à des dates relativement espacées. Les informations ont alors de fortes chances d'être obsolètes. Les mises à jour sont manuelles et donc longues et coûteuses.

Aux éditions LAROUSSE, la question de la mise à jour se pose depuis longtemps. Plusieurs stratégies ont été envisagées, la stratégie actuelle consistant à demander aux auteurs de mettre le moins possible d'informations susceptibles d'évolutions. Une stratégie intéressante a également été mise en place : il était demandé aux auteurs de se prononcer au moment de la rédaction des articles sur la nature potentiellement évolutive des informations (par balisage XML).

[5] `http://fr.encarta.msn.com/`

[6] `http://www.encyclopedie-gratuite.fr/`

[7] Le packaging concerne la mise en page des livres (intégration des textes, d'images, de graphisme, etc.).

[8] ou *collection vendue à tempérament*

Mais cette solution a été abandonnée car elle s'est avérée contraignante pour les auteurs (temps supplémentaire de rédaction, difficulté d'évaluer la nature de l'information). Aujourd'hui, les entrées des dictionnaires sont mises à jour par un processus entièrement manuel, principalement en fonction de l'actualité et des événements importants. La stratégie est identique pour les versions papier et les versions en ligne des encyclopédies.

Il existe paralèllement les solutions dites coopératives : Wikipédia est naturellement cité en exemple. La solution coopérative consiste à laisser les internautes participer eux-mêmes à la rédaction des articles en ligne et ils participent donc tout naturellement au processus de mise à jour des documents. Cependant la réactivité des internautes n'est pas toujours rapide, elle dépend pour beaucoup des sujets traités et les informations obsolètes peuvent rester longtemps [1].

Dans tous les cas, la possibilité d'une automatisation de l'aide à la mise à jour de documents apparait comme centrale pour un secteur qui doit être capable de satisfaire une forte demande en terme de validité de l'information.

Nous proposons un prototype de navigation intra-documentaire signalant aux rédacteurs le caractère potentiellement obsolescent des informations situées dans les parties d'un document encyclopédique.

3 Définition du problème : repérer les zones à mettre à jour

Rechercher de manière automatique les segments textuels susceptibles de contenir de l'information à mettre à jour est loin d'être une tâche triviale. Nous décrivons dans cette section les principaux travaux en lien avec cette problématique (section 3.1), puis nous présentons dans la section 3.2 la notion d'obsolescence.

3.1 Rechercher l'expression du changement ou de l'évolution

À notre connaissance, il n'existe pas de travaux portant sur la problématique de mise à jour de l'information. Le domaine avec lequel nous avons le plus de points communs à la fois dans les objectifs et dans les méthodes mises en oeuvre est la veille stratégique [9].

La veille stratégique est définie comme l'activité mettant en œuvre des techniques d'analyse d'informations sur un produit ou un procédé et sur l'état de l'art et l'évolution de son environnement scientifique, technique, industriel ou commercial. L'objectif est de collecter, organiser, puis analyser et diffuser les informations pertinentes qui vont permettre d'anticiper les évolutions et qui vont aider à l'innovation. Techniquement, la veille stratégique fait principalement appel aux techniques de recherche d'information.

L'une des techniques consiste à rechercher les indicateurs de nouveauté et d'innovation : ce type d'informations permet à une entreprise de rester au fait des innovations technologiques. Un certain nombre de travaux ont été menés sur la recherche de la nouveauté ou encore sur la question de la variation terminologique.

Lors de l'évaluation TREC en 2004, la "novelty track"[10] avait pour objectif de tester les capacités des systèmes à repérer automatiquement des informations nouvelles dans les textes (`http://trec.nist.gov/tracks.html`).

Dans un cadre de recherche similaire, le travail de F. Ibekwe-SanJuan [2] a pour objectif le repérage d'indices de nouveauté pour détecter ce qui change dans le contenu des textes eux-mêmes. L'auteur met en œuvre une méthodologie dans laquelle nous nous retrouvons notamment parce qu'elle fait appel à des indices linguistiques. L'auteur s'inspire des travaux de S. Teufel, J. Carletta et M. Moens [3]. Elle travaille sur des textes courts (titres et résumés scientifiques) et

[9] Ou veille scientifique ou veille technologique.

[10] Un résumé des résultats est donné ici : `http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf`.

oriente ses recherches sur l'expression de l'apport de l'auteur. Elle met en relief trois types d'indices : ceux qui rendent compte des objectifs, ceux qui montrent les contributions et résultats et enfin ceux qui apportent des conclusions. Ses travaux sont menés sur des textes en langue anglaise et les indices textuels qu'elle met en évidence sont des expressions du type « Here, we propose a novel (...) approach », « we discuss recent developments » (information apportée : la nouveauté), « Our research suggests that », « Results confirm that » (information apportée : résultats/ contribution/ conclusion), « In this article/ paper/ study, we examine/ investigate/ describe » (information apportée : objectif). À l'issue d'une observation manuelle des indices, elle crée des automates qui dans un nouveau corpus vont repérer automatiquement les types d'indices sus-cités. Les résultats sont encourageants. L'auteur suppose que les indices qu'elle met en lumière sont généralisables à des domaines scientifiques différents moyennant de légères variations. Elle reste malgré tout prudente quant au fait que toute nouveauté n'est pas systématiquement encadrée, marquée par des indices textuels et qu'il serait probablement pertinent de les coupler à des indices fréquentiels et temporels.

Les travaux de A. Condamines, J. Rebeyrolles et A. Soubeille [4] concernent la question de l'évolution des connaissances à travers l'étude des termes au sein de domaines spécifiques (l'aérospatial). L'objectif est *« d'identifier et décrire les formes privilégiées du changement sémantique à partir de l'analyse linguistique d'un corpus spécialisé construit de façon à rendre possible l'observation d'évolutions de connaissances »*. Ces auteurs utilisent une méthode *outillée* par des traitements automatiques (extraction de termes, étiquetage grammatical, concordancier). Leur façon d'utiliser les corpus diffère cependant de la nôtre puisque leur choix a consisté à comparer les « mêmes » textes pris à des moments différents de leur évolution ce qui permet une réelle étude diachronique alors que notre étude est exclusivement synchronique (collecte de textes à un moment donné). La thèse d'A. Picton [5] va également dans ce sens : l'auteur recherche les indices linguistiques permettant d'accéder à l'évolution des connaissances sur des périodes courtes. Sur la base de l'étude de quatre types d'indices (fréquences, contextes d'évolution, variantes et dépendances syntaxiques), elle cherche à associer un ou plusieurs indices avec un ou plusieurs types d'évolution spécifique.

En ce qui concerne notre tâche, considérer la tâche de mise à jour uniquement en fonction du repérage d'indices de nouveauté est clairement insuffisant. Nous proposons de rechercher ce que nous nommons des « segments d'obsolescence ». La section suivante définit cette notion.

3.2 L'obsolescence : quelle réalité pour l'édition ?

Le mot « obsolescence » vient du latin « tomber en désuétude ». La définition de ce terme par le Grand Robert s'inscrit dans le domaine technique :

> *« Vieillissement de l'équipement industriel, dû à l'apparition d'un matériel nouveau » (Le Grand Robert)*

Les aspects économiques et sociétaux sont intrinsèquement liés à l'obsolescence ainsi que le montre la citation suivante :

> *« L'obsolescence a été étudiée et changée en technique. Les spécialistes de l'obsolescence connaissent l'espérance de vie des choses : trois ans, une salle de bain ; cinq ans, un living-room ; huit ans, un élément de chambre à coucher ; trois ans, l'aménagement d'un point de vente local, une auto, etc. »* (Henri Lefebvre, La vie quotidienne dans le monde moderne, in Le Grand Robert, p. 157)

Dans un sens, les encyclopédies actuelles répondent à cette définition : le client d'une encyclopédie (ou d'un dictionnaire) peut trouver sa collection obsolète car il n'y trouve pas les derniers mots, concepts et définitions à la mode ou nouvellement créés et utilisés. Il cherchera alors naturellement à se procurer un exemplaire plus récent, plus en adéquation avec ses attentes : chiffres et dates récentes, nouveaux mots, derniers événements politiques, économiques, sociaux, etc.

Parallèlement, une encyclopédie est un objet qui a également pour but de permettre l'accès à une forme de savoir absolu et vrai pour tout citoyen. Pour rappel, les premières lignes écrites par Diderot en 1751 dans l'Encyclopédie :

« Le but d'une encyclopédie est de rassembler les connaissances éparses sur la surface de la terre ; d'en exposer le système général aux hommes avec qui nous vivons, et de le transmettre aux hommes qui viendront après nous ; afin que les travaux des siècles passés n'aient pas été inutiles pour les siècles qui succéderont ; que nos neveux devenant plus instruits, deviennent en même temps plus vertueux et plus heureux ; et que nous ne mourions pas sans avoir bien mérité du genre humain. »

Dans une encyclopédie, suggérer l'existence de zones contenant de l'information obsolète ne remet pas en cause le bien-fondé de l'existence même de l'encyclopédie : il est en revanche indispensable de savoir le repérer et le corriger, si le besoin se fait sentir, afin de parvenir à un juste équilibre entre des informations anciennes mais nécessaires à la compréhension du monde d'aujourd'hui et des informations qui ont pu évoluer et se modifier et qui sont susceptibles de rendre obsolète un tel objet (aux sens économique et industriel du terme).

Mais comment définir la validité ou la véracité d'une information ? Un article paru dans le journal Nature en 2005 compare des articles scientifiques tirés de l'encyclopédie collaborative Wikipédia à des articles de l'encyclopédie Britannica. Si l'étude conclut sur une fiabilité de l'information à peu près égale dans les deux sources, elle met surtout le doigt sur la difficulté à mettre au point des grilles d'évaluation objectives : l'expérience a été menée sur des textes scientifiques pointus et où les différences de point de vue sont moindres et certains articles ont même été coupés.

Une connaissance est difficilement évaluable en termes de vérité absolue ou non (vrai ou faux) même si cette connaissance porte sur un fait vérifiable du monde. Il semble plus prudent de penser les connaissances d'une manière graduelle en considérant qu'elle sont, à un bout de l'échelle, certaines, et à l'autre bout de l'échelle, peu certaines. Entre ces deux extrêmes, une connaissance factuelle peut être soumise au doute, être jugée acceptable en attendant la preuve du contraire.

Dans ce contexte, nous considérons qu'une information obsolescente est une information qui n'est potentiellement plus vraie, qui peut être jugée douteuse ou subjective ou encore qui n'est plus *valable* à $T + n$ du fait de l'évolution temporelle ou technique. n correspond à une période temporelle définie arbitrairement : pour ce travail, nous considérons $T + 1\ an$ au minimum, étant donné que les maisons d'édition proposent généralement des versions réactualisées tous les ans.

3.3 Le segment d'obsolescence : une unité fonctionnelle et textuelle

De la présentation de notre projet découle l'idée qu'un segment d'obsolescence se définit d'abord par rapport à un usage concret, un besoin réel, à savoir la mise à jour éditoriale. Un segment d'obsolescence présente la particularité majeure de contenir une information dont la caractéristique est d'être susceptible d'évolution dans le temps. Ce segment textuel peut également être pertinent parce qu'il véhicule des connaissances qui, relativement à des besoins éditoriaux, nécessiteraient d'être réactualisées.

L'exemple 1 présente un segment d'obsolescence. L'auteur y exprime une issue possible et probable concernant les recherches sur le sida.

La fiche d'où est extrait ce passage a été éditée et distribuée dans le courant de l'année 2003 par les Éditions Atlas. Ce type de fiche est destiné à être distribué dans le cadre d'éditions au long cours : un client qui s'abonne à l'encyclopédie en 2007 est susceptible de recevoir cette fiche, écrite en 2003 et dans laquelle cet exemple apparaît. Cependant, pendant ce laps de temps (de 2003 à 2007), soit certaines des prédictions formulées par l'auteur se seront réalisées, soit elles auront été repoussées par les scientifiques ou encore de nouvelles données peuvent entrer en jeu. Il est donc tout à fait souhaitable que ce segment de texte ait été préalablement mis à jour.

Dans cet exemple, nous observons l'importance de la composante temporelle. L'objectif de cette étude n'est ni de procéder à des calculs de la référence

1. Actualité

§ Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux. Toutefois, il convient de rappeler un certain nombre de découvertes très récentes. En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

1.1. Un vaccin contre le sida ?

§ Des recherches portant sur les prostituées [...]. La recherche se tourne justement aujourd'hui vers des vaccins qui [...]. Des expériences ont été faites pour [...]. En juin 2003, une équipe de biologistes américains a obtenu des résultats qui pourraient laisser envisager [...]. Les chercheurs sont parvenus [...]. Cette découverte pourrait aboutir à la mise au point d'un antigène [...]. [. . .]

Source : Corpus ATLAS (fiche Médecine - Le Sida)

Exemple 1 – Un segment à mettre à jour

temporelle, ni de chercher à associer un événement particulier à une date particulière comme c'est le cas en extraction d'information ou dans les systèmes de question-réponse. Nous ne cherchons pas non plus à valider le fait que l'événement « une équipe de biologistes américains a obtenu des résultats qui [...] » est vrai, ni qu'il s'est réellement produit en « juin 2003 ».

Nous recherchons les segments pour lesquels il est pertinent de penser que l'information donnée est susceptible d'avoir évolué entre le moment de l'édition de la fiche et le moment de sa lecture potentielle par un client. Dans l'exemple 1, la phrase « Cette découverte pourrait aboutir à la mise au point d'un antigène » doit être vérifiée et mise à jour. Le type d'information recherché peut être local (une date, un chiffre, un nom de société, etc.) ou au contraire à granularité variable, de la taille de la phrase à celle du paragraphe tout entier.

Pour comprendre et étudier ce phénomène à garnde échelle, nous avons constitué un corpus, le corpus ENCYCLO, réunissant des articles extraits des trois encyclopédies différentes :

– les mémofiches publiées aux éditions Atlas ;
– des articles extraits du Grand Universel Larousse ;
– des articles extraits du Grand Larousse Informatisé.

Ce corpus compte environ 282 000 mots dans 89 articles encyclopédiques différents et traitant de sujets divers, de l'histoire à la médecine en passant par la géographie ou encore les sciences et techniques.

3.3.1 Quelle réalité dans les encyclopédies

Le corpus ENCYCLO a été annoté manuellement selon le caractère obsolescent ou non des segments[11] qui le composent : un expert linguiste[12] a annoté le corpus ATLAS et le corpus LAROUSSE qui a été également annoté par trois experts rédacteurs[13]. Comme le montrent les chiffres du tableau 1, la part des segments obsolescents est de 15 % environ dans le corpus [ENCYCLO].

	ATLAS	LAROUSSE	ENCYCLO
nombre total de phrases	7142	2874	9916
nombre de phrases obsolescentes	927	581	1508
pourcentage de phrases obsolescentes	12,9 %	20,2 %	15,2 %

TAB. 1 – Proportion de segments obsolescents dans notre corpus ENCYCLO

La différence de proportion d'obsolescence entre le sous-corpus ATLAS et le sous-corpus LAROUSSE vient du fait que les textes du sous-corpus ATLAS ont été réunis selon les données rendues disponibles par l'éditeur : il contient proportionnellement plus de fiches appartenant au domaine de la géographie, domaine qui se trouve être également très enclin au besoin de mise à jour (en géographie, 27,7 % des phrases sont potentiellement à mettre à jour alors qu'en histoire, leur proportion n'est que de 8,1 %). Le sous-corpus LAROUSSE est quant à lui plus homogène.

Une première analyse des segments annotés manuellement nous amène à distinguer deux grandes classes d'obsolescence : d'un côté les segments dont l'information est devenue fausse (la connaissance est envisagée dans un moment T ; à $T+1$, elle est susceptible d'avoir évolué) ; de l'autre, ceux, dont l'information n'est plus pertinente (d'un point de vue informationnel) au moment où elle est lue. Comparons les deux exemples construits (et donc simplifiés) suivants :

(1) *Aujourd'hui, le PIB par habitant de la France est de 27 600 dollars.*

(2) *En 2004, le PIB par habitant de la France est de 27 600 euros.*

Dans l'exemple (1), sachant qu'on est actuellement en 2010, et que le lecteur va naturellement interpréter l'adverbial « aujourd'hui » comme étant l'année en cours, l'information est fausse puisque le PIB de la France le plus actuel (chiffres de 2008) est de 33 800 dollars[14].

À l'inverse, l'exemple (2) montre un cas où l'information restera toujours vraie : en 2004, le PIB par habitant de la France sera toujours de 27 600 dollars. Une mise à jour éditoriale sera cependant nécessaire si l'objectif du rédacteur est de fournir les résultats les plus récents par rapport à la date en cours : il faudra donc vraisemblablement actualiser à la fois la référence temporelle (« En 2009 ») et la valeur du PIB associée.

Cette distinction montre que l'obsolescence est un phénomène complexe qui ne saurait se réduire à un traitement simple.

[11] L'unité prise en compte est la phrase.

[12] Moi-même, Marion Laignelet.

[13] Rédacteurs de la société Larousse.

[14] La situation serait sans doute identique avec une phrase ne contenant pas d'adverbial temporel. En effet, lorsqu'il n'y a pas de référence temporelle précise, le temps verbal donne certaines indications : ici le présent suggère une interprétation déictique. Mais dans les corpus ce n'est pas toujours le cas (présent historique par exemple).

3.3.2 L'obsolescence, un phénomène consensuel

Parce que nous disposons d'une multi-annotation humaine[15], il nous a été possible d'évaluer le taux d'accord de jugement sur l'obsolescence. Nous avons mesuré les taux de recouvrement des annotations manuelles. Il en ressort que le sous-corpus LAROUSSE est en moyenne composé de 10 à 15 % de segments obsolètes par annotateur et que l'*accord observé* entre chacun de ces juges se situe entre 87 et 92 %.

Le cœfficient Kappa est traditionnellement utilisé pour évaluer les degrés d'accord entre juges. Concernant l'obsolescence, le taux d'accord est situé entre 0.35 et 0.50 : ce score très faible est directement lié à la forte disproportion des classes (15 % de segments obsolescents contre 85 % de segments non obsolescents). L'accord entre nos juges est mieux traduit par le cœfficient *r* de Finn[16] [6].

Ce cœfficient permet d'aplanir la disproportion des classes en comparant la proportion des accords observés à une situation aléatoire considérant, dans notre cas bien précis, que chaque annotateur a une chance sur deux de déclarer un segment obsolescent (en situation de hasard). Les scores pour le cœfficient *r* de Finn varient de 0.75 pour l'accord le plus bas (les codeurs 2 et 4) à 0.83 pour l'accord le plus haut (codeurs 1 et 3).

Ces résultats montrent qu'il n'y a pas une grande variation de jugement entre les quatre experts sur la nature obsolescente ou non d'un segment. En d'autres termes, cela nous conforte dans l'idée que l'obsolescence est un phénomène qui fait suffisamment consensus pour être automatisé. Mais cela montre également qu'il s'agit d'un phénomène difficile à appréhender et que, dans tous les cas, il serait illusoire de penser qu'on pourra mettre au point un prototype idéal qui fera mieux que l'humain.

3.3.3 L'obsolescence, un phénomène dépendant de la thématique

L'annotation manuelle de l'obsolescence met en évidence la forte variation du nombre de segments selon les rubriques thématiques. Le tableau 2 récapitule le nombre et le pourcentage de segments obsolescents en fonction des rubriques traditionnellement présentes dans une encyclopédie.

Dans ce tableau, on observe que le nombre de segments obsolescents varie selon la thématique : de 0 %[17] dans les textes traitant d'*Art et Littératures* à presque 30 % pour la *Géographie*.

4 Un outil pour repérer les segments potentiellement obsolescents dans les encyclopédies

Le dispositif expérimental que nous avons mis en place pour rendre compte du phénomène de l'obsolescence dans les textes encyclopédiques se décompose en trois parties principales [7, 8].

[15] Le sous-corpus LAROUSSE a été annoté par quatre experts différents.

[16] Nous utilisons l'algorithme existant dans le logiciel R.

[17] Ce qui ne veut pas dire qu'il n'y aura jamais de mise à jour dans les fiches relevant de ce domaine ; ces chiffres rendent compte de la réalité de notre corpus, non pas de la réalité du type encyclopédique en général.

	nombre total de segments	nombre de segments obsolescents	Pourcentage de segments obsolescents
Géographie	1816	503	**27.7 %**
Économie/ Société/ Droit	1916	290	15.1 %
Sciences et Techniques/ Faune et Flore	1527	297	19.5 %
Histoire	1513	123	8.1 %
Sport	525	26	19 %
Arts et Littératures	332	0	**0 %**
Médecine	1720	123	7.1 %
rubrique inconnue	567	146	**25.7 %**
TOTAL	9916	1508	15.2 %

TAB. 2 – Les segments obsolescents selon les rubriques de l'encyclopédie

La première étape consiste à repérer dans notre corpus annoté manuellement (le corpus [ENCYCLO]), les expressions linguistiques susceptibles d'être de bons indices de l'obsolescence : c'est le rôle de l'outil ALIDIS (**A**nnotation **LI**nguistique des **DIS**cours) développé à l'aide de la plateforme de développement de TAL LinguaStream [9]..

Le tableau 3 résume l'ensemble des traits textuels repérés de manière automatique par notre outil. Il rend compte de la variété des types de traits pris en compte et de leur variation en termes de granularité dans le document.

Dans un second temps, à partir des résultats fournis par ALIDIS, une étape de transformation des données textuelles (en XML) en un format matriciel permet un traitement statistique des indices linguistiques et discursifs. Cette transformation est basée sur un modèle pivot (*i.e.* un modèle conceptuel des données) : c'est l'objectif de l'outil OCAS (**O**util de **C**réation d'**A**bstraction **S**émantique) [7].

Enfin, les données générées par OCAS font leur entrée dans l'outil STAAT qui produit une analyse statistique. Nous avons mené trois types d'analyse statistique : un module de statistiques descriptives basiques (corrélation), une Analyse en Composantes Principales (ACP) et enfin un module d'apprentissage automatique

Types de marqueurs	**Structure de traits**		**Exemples**
	Trait 1	**Trait 2**	
Adverbiaux temporels	nature : *ponctuel, inachevee, deictique, duree, iteration*	sitTps : *anteriorite++, anteriorite, coincidence, posteriorite, indetermine*	pendant dix ans $\left[\begin{array}{l} nature : durée \\ sitTps : indéterminé \end{array}\right]$
Temps verbaux	temps : *passeComp, passeAnt, plusQuePft, futAnt, condPasse, present, passeSimple, imparfait, futur, conditionnel*		Les 3e et 4e tranches [...] [sont]] $\left[temps : présent \right]$ en cours de réalisation.
Périphrases verbales	accomplissement : *debut, fin, deroulement, continuite*		Les 3e et 4e tranches [...] [sont en cours de] $\left[accomplissement : déroulement \right]$ réalisation.
Entités nommées	classe : *personne, lieu, sigle, web, mail, marque, geopolitique, mesure*	sousClasse : *riviere, pays, ville, evolutif, fixe,...*	Les expéditions d'Arzila [à Tanger] $\left[\begin{array}{l} classe : lieu \\ sousClasse : ville \end{array}\right]$; supérieurs à [7 600 euros] $\left[\begin{array}{l} classe : mesure \\ sousClasse : évolutif \end{array}\right]$
Expressions du point de vue	type : *distance, jugement, recence, prevision, importance, jugePerso, source, thematique, restriction*		Il s'agit d'[un véritable enjeu] $\left[type : importance \right]$; Selon le rapport de l'INSEE $\left[type : source \right]$
Argumentatifs	type : *correction, explication, opposition, conséquence, temporelle, exemplification,...*		[Mais] $\left[\begin{array}{l} classe : argum \\ sousClasse : opposition \end{array}\right]$
Type de phrase	type : *exclamation, assertion, interrogation*		
Position de l'indice dans la phrase	position : *debut, fin, amorce*		
Position de la phrase dans le paragraphe	position : *debutParag, finParag, seuleInParagraphe*		
Position du paragraphe dans le document	position : *debutZone, finZone, debutDivision, finDivision*		

TAB. 3 – Résumé des marqueurs textuels et discursifs repérés automatiquement

exploitant un système à base de règles d'association [8].

Nous avons cherché à rendre cette méthodologie reproductible et adaptable pour d'autres tâches. Nous souhaitons pouvoir modifier aisément les indices linguistiques à prendre en compte, les annotations manuelles mais également les corpus en entrée du dispositif ou encore les outils statistiques en sortie.

5 Résultats et performances

Pour pouvoir juger des résultats, nous avons mis en place deux systèmes de base. Ils constituent également une première méthode naïve de classification. Ces systèmes s'organisent ainsi :

Système 1 : partant du constat empirique que les traits les plus corrélés à l'obsolescence sont des valeurs numériques et des dates, nous testons tout d'abord si la simple présence de chiffres dans une phrase est suffisante pour déterminer sa nature obsolescente ou non ;

Système 2 : la seconde méthode de base est la présence d'au moins un des traits les plus corrélés à la variable *obsol* : expressions temporelles déictiques, ponctuelles ou de durée lorsqu'elles réfèrent à une date proche du moment d'énonciation ou située dans le futur, les temps/modes futur et conditionnel, les adverbiaux exprimant un point de vue de type récence (« les territoires actuels ») ou prévision (« les recherches à venir »).

Ces deux systèmes ont été créés à partir d'expériences au cours desquelles nous avons mesuré la corrélation de la variable *obsol* (interprétée à partir des annotations manuelles de l'obsolescence par les experts) avec chacune des autres variables de la base (les traits repérés de manière automatique par l'outil ALIDIS). Nous avons pour cela utilisé le logiciel SPAD[18].

Voici la performance de tels systèmes pour le repérage des segments obsolescents comparée aux résultats de notre prototype basé sur un système d'apprentissage automatique à base de règles d'association[19]. Ce prototype exploite l'ensemble des traits linguistiques présentés dans la section 4.

	Précision	Rappel	F-Score
Base 1	23	31	26
Base 2	30	39	37
corpusComplet	32.9	78.8	46.4

TAB. 4 – Comparaison des performances du classifieur selon les différentes vues sur le corpus d'apprentissage

On constate que le prototype exploitant les traits linguistiques hétérogènes a de meilleures performances que des systèmes plus "naïfs". Même s'ils ne sont pas suffisants pour une industrialisation, ces résultats sont encourageants si l'on considère qu'il s'agit là d'une étude exploratoire sur la question de l'obsolescence dans les encyclopédies.

De plus, nous voulons un système qui privilégie le rappel par rapport à la précision : en effet, oublier une révision est plus grave qu'indiquer inutilement un segment à réviser et donc, il vaut mieux repérer trop de segments même s'ils ne sont pas obsolescents.

Une observation plus approfondie des données fait ressortir de grandes différences dans les résultats lorsqu'on prned en compte les rubriques thématiques. Le tableau suivant indique les performances de notre prototype selon les thématiques des encyclopédies.

[18] `http://www.coheris.fr/fr/page/home.html`

[19] Méthode de classification supervisée en 10-cross-validation.

	Précision	Rappel	F-Score
Sport	0,33	0,33	0,33
Économie	0,65	0,52	**0,58**
Médecine	0,30	0,43	0,35
Histoire	0,01	0,13	0,03
Société	0,10	0,33	0,15
Géographie	0,69	0,84	**0,76**
Biologie	0,07	0,86	0,13
Arts et littératures	0,45	0,65	**0,54**
Divers	0,37	0,65	0,47

TAB. 5 – Performances du prototype en fonction des rubriques thématiques

Les chiffres de ce tableau montrent que notre classifieur a de meilleurs résultats avec les textes appartenant au domaine de la géographie et de l'économie. Nous supposons que cela est lié au fait qu'il y a plus de segments obsolescents dans ce type de domaines (*cf.* le tableau de la section 3.3.3) que dans les autres ; l'apprentissage se faisant sur un nombre de cas plus important, cela engendrerait une meilleure représentativité de ce domaine.

L'importance du corpus est également mise en avant : non seulement, il est important de mettre en place un corpus représentatif pour la tâche visée (le type encyclopédique est particulier car il rend compte de connaissances thématiques diverses) mais qui soit également à même de rendre compte des phénomènes spécifiques visés (dans notre cas, l'obsolescence ne représente que 15 % des phrases du corpus).

Concernant l'importance de la thématique, l'Analyse en Composantes Principales (ACP)[20] qui a été faite sur nos données [7] souligne des regroupements intéressants de traits textuels particuliers par rapport aux thématiques.

Les résultats de l'ACP font ressortir que dans les 15 premiers facteurs, des oppositions systématiques sont établies en fonction des différentes rubriques : Économie/Histoire/Géographie sont opposés à Médecine dans le facteur 1 ; Géographie s'oppose à Histoire dans le facteur 2 ; Économie à Sport dans le facteur 3, Sciences et Techniques à Géographie dans le facteur 7, Sport à Histoire/Art et Littératures dans le facteur 9, etc.

Par exemple, le facteur 7 de notre ACP, illustré à travers l'exemple 2, met en évidence les phrases constituées d'expressions temporelles de type *ponctuel coïncidence* (V028) en position normale et d'amorce (V106), d'entités nommées de type *sigle* (V020) et d'expressions du point de vue de type *source* (V081). La rubrique Sciences et Techniques (V099) est fortement corrélée à ces variables. Ces phrases ont une probabilité forte d'être obsolescentes.

[20]Nous avons utilisé le logiciel SPAD. L'analyse menée est normée, c'est-à-dire qu'on ne cherche pas à donner plus d'importance aux phrases qui contiennent beaucoup d'indices. Toutes les variables quantitatives (continues) sont actives, soit 146 variables. La variable *obsol* est traitée comme une variable *active* au même titre que toutes les autres variables car ce sont l'ensemble des relations (corrélations) entre *obsol et* les autres variables qui nous préoccupent.

D'après le BIT, dans les pays de l'OCDE, 37 % des travailleurs étaient affiliés à un syndicat en 1975 ; ils n'étaient plus que 28 % en 1988 et l'érosion des effectifs s'est poursuivie au cours de la décennie suivante (taux aujourd'hui souvent inférieur à 20 %), notamment en lien avec le recul de l'identité ouvrière. [. . .]

n° d'individu : 1237458232636

Exemple 2 – Les expressions temporelles de type *ponctuel coïncidence*, les entités nommées de type *sigle* et l'expression du point de vue de type *source* dans un segment obsolescent (individu représenté sur l'axe 7 - positif)

De l'autre côté de cet axe 7, une phrase composée de temps verbaux au passé simple (V143), d'entités nommées de type *lieu pointCardinal* (V095) et de type *lieu ville/rivière/pays* (V139, V083 et V045) ne sera probablement pas à mettre à jour. Ces corrélations semblent fréquemment associées aux phrases issues de textes de Géographie (V094).

6 Conclusion

L'objectif de ce travail est de chercher et de proposer des méthodes et techniques pour produire un repérage des segments textuels contenant une information potentiellement obsolescente. Il ne s'agit pas de se substituer aux rédacteurs des encyclopédies en effectuant à leur place la mise à jour des informations. Nous proposons le développement d'outils s'intégrant dans un système d'aide à la mise à jour des textes encyclopédiques.

Les résultats présentés dans ce papier nous amènent à redéfinir, avec l'aide des rédacteurs d'encyclopédie, la notion d'obsolescence dans le cadre d'une tâche de mise à jour. Cette redéfinition des besoins est centrale et nécessite la participation entière des utilisateurs finaux d'un tel outil. Ce sont eux qui pourront permettre la mise en oeuvre d'un protocole d'annotation manuelle de l'obsolescence qui soit plus homogène, plus précis, plus fiable et qui permettra vraisemblablement d'améliorer les résultats de notre prototype.

Les marqueurs linguistiques doivent également être affinés. S'il est évident que la diversité des indices que nous prenons en compte est un atout, la sémantique qui leur est attachée ne l'est pas forcément dans tous les cas. Par exemple, les indices concernant le point de vue du locuteur doivent être ré-évalués et organisés d'une manière plus explicite et en accord avec des théories linguistiques précises, ce qui n'est pas réellement le cas à l'heure actuelle. Ils doivent également être pensés et organisés en fonction des thématiques des encyclopédies. Les traits textuels pertinents pour la géographie ne le seront pas systématiquement pour l'histoire.

La technique mise en place repose sur l'utilisation de connaissances linguistiques hétérogènes. D'une manière générale, la méthodologie que nous proposons est reproductible et réutilisable. Une autre perspective consiste à tester notre méthodologie sur d'autres textes : sur des textes extraits de Wikipédia afin de valider nos résultats sur une autre source de données encyclopédiques, sur des textes différents comme des manuels scolaires ou des rapports techniques ou d'entreprise.

Références

[1] P. Gourdain, F. O'Kelly, B. Roman-Amat, D. Soulas, and T. vonDrosteHülshoff. *La Révolution Wikipédia – Les encyclopédies vont-elles mourir ?* Mille et une nuits, 2007.

[2] F. Ibekwe-SanJuan. Annotation d'indices de nouveautés dans les écrits scientifiques et techniques. In *Colloque Indice, Index, Indexation*, 2005.

[3] S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, 1999.

[4] A. Condamines, J. Rebeyrolle, and A. Soubeille. Variation de la terminologie dans le temps : une méthode linguistique pour mesurer l'évolution de la connaissance en corpus. In *Actes Euralex International congress*, pages 547–557, Lorient, France, 2004.

[5] A. Picton. Combining clues to explore knowledge evolution. In *Actes de la Conférence internationale Terminology and Knowledge Engineering (TKE)*, Copenhague, Danemark, 2008.

[6] G. Hripcsak and D.F. Heitjan. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35(2) :99–110, 2002.

[7] M. Laignelet. *Analyse discursive pour le repérage automatique de segments obsolescents dans les documents encyclopédiques*. PhD thesis, Université de Toulouse - Le Mirail, 2009.

[8] M. Laignelet and F. Rioult. Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009*, 2009. prix du "Meilleur papier".

[9] A. Widlöcher and F. Bilhaut. La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)*, Dourdan, France, 2005.