RÉCITAL 2002, Nancy, 24–27 juin 2002

Normalisation de documents par analyse du contenu à l'aide d'un modèle sémantique et d'un générateur

Aurélien Max
Groupe d'Etude pour la Traduction Automatique (GETA CLIPS-IMAG)
Xerox Research Centre Europe (XRCE)
Grenoble, France
aurelien.max@imag.fr
date de soutenance prévue : Novembre 2003

Mots-clefs – Keywords

analyse de contenu, génération, création assistée de documents, normalisation de document
content analysis, generation, document authoring, document normalization

Résumé - Abstract

La problématique de la *normalisation de documents* est introduite et illustrée par des exemples issus de notices pharmaceutiques. Un paradigme pour l'analyse du contenu des documents est proposé. Ce paradigme se base sur la spécification formelle de la sémantique des documents et utilise une notion de similarité floue entre les prédictions textuelles d'un générateur de texte et le texte du document à analyser. Une implémentation initiale du paradigme est présentée.

This paper discusses *document normalization* and gives examples based on a class of pharmaceutical documents. The discussion is based on a paradigm for document content analysis. This paradigm focusses on a formal specification of document semantics and uses a fuzzy matching measure between the textual predictions of a natural language generator and the input document. An initial implementation is presented.

1 Introduction

Nous introduisons la problématique de la *normalisation de documents*, et nous l'illustrons sur des exemples issus de notices pharmaceutiques. La normalisation d'un document est définie comme la reconnaissance des *buts communicatifs prédéfinis* exprimé dans celui-ci, puis leur reformulation par un processus de génération contrôlée. Dans cet article, nous définissons des besoins sur la spécification du contenu des documents pour leur normalisation, puis nous montrons comment des modèles existants développés pour la création assistée de documents multilingues peuvent être utilisés. Nous proposons ensuite un paradigme pour l'analyse du contenu qui utilise de tels modèles en conjonction avec la notion de *génération inversée floue*, et nous montrons comment le contenu d'un document source peut être retrouvé heuristiquement par le biais d'une procédure de recherche admissible. Dans un dernier temps, nous présentons une implémentation initiale d'un système de normalisation de documents existants.

2 Normalisation de documents

2.1 Etude de cas: les notices pharmaceutiques

Afin de motiver la tâche de normalisation de documents, nous avons étudié un corpus contenant 50 notices pharmaceutiques en anglais pour des médicaments analgésiques.[1] Les notices, issues de différentes pharmacies en ligne en Grande-Bretagne et aux Etats-Unis, constituent une collection hétérogène de documents comparables pour des médicaments similaires ayant l'aspirine comme principal principe actif. Notre étude de corpus a révélé différents types de variation, que nous allons illustrer sur deux exemples de notices (voir figure 1), ANA (pour ANADIN) et ALK (pour ALKA-SELTZER).[2]

Tout d'abord, les structures des deux notices sont différentes, ce qui se traduit par exemple dans la façon dont les informations relatives aux avertissements (*warnings*) sont organisées. ANA distingue une section d'avertissements assez générale (*Warnings*), une section sur les effets indésirables possibles (*Side effects*), alors que ALK a une section sur les intéractions médicamenteuses possibles (*Drug interaction precautions*), une section d'avertissements (*Warnings*), et une section sur les avertissements propres à la consommation d'alcool (*Alcohol warning*). ANA donne les informations relatives aux intéractions médicamenteuses possibles dans sa section sur les avertissements (*Warnings: You should ask your doctor before taking aspirin if you are taking medicines for...*). A l'inverse, les effets indésirables possibles, qui apparaissent dans la section générale sur les avertissements dans ALK (*If ringing in the ears or a loss of hearing occurs...*), ont une section propre dans ANA.

Certains types de contenu peuvent être exprimés ou non, ce qui reflète des décisions prises par l'organisme responsable de la diffusion de la notice. Ainsi, ALK a une section specifique sur les avertissements liés à la consommation d'alcool (*Alcohol warning*); l'effet indésirable possible correspondant (hémorragie digestive haute, *stomach bleeding*) est également exprimé dans ANA dans sa section sur les effets indésirables possibles (*Side effects*), mais sans référence

[1] Le choix de la langue du corpus a principalement été motivé par la disponibilité de documents électroniques de différentes origines en anglais. L'approche que nous présentons dans cet article s'applique également au français.

[2] Ces deux médicaments contiennent également d'autres principes actifs, ce qui a bien entendu des répercussions sur le contenu de leur notice respective.

ANADIN (source: www.pharmacy2u.co.uk)
Indications For the symptomatic relief of influenza and common colds. Also indicated for the treatment of mild to moderate pain, including headache, migraine, neuralgia, dental pain, period pain and muscular aches and pains.
Directions ...
Ingredients Capsule containing: Aspirin (acetylsalicylic acid) 500 mg ...
Warnings Do not take aspirin if you are allergic to aspirin or to any other ingredients, have a stomach ulcer or have haemophilia. Avoid in asthmatics and pregnancy particularly the final trimester. Aspirin has also been associated with increased risk of Reye's Syndrome when given to children with a fever. For this reason the use of aspirin is not recommended in children under 12 years of age. You should ask your doctor before taking aspirin if you are taking medicines for blood clots (thrombosis) such as warfarin or gout such as probenecid. ...
Side effects In people sensitive to aspirin, reactions such as asthma attacks and skin rashes may occur occasionally. Aspirin may induce gastric irritation, nausea, dyspepsia and stomach bleeding. Consult your doctor or pharmacist if you have any side effects after taking this product. ...

ALKA-SELTZER (source: www.drugstore.com)
Indications For fast relief of heartburn, acid indigestion, sour stomach with headache or body aches and pains. ... Effective for pain relief alone; headache or body and muscular aches and pains.
Directions ...
Ingredients Active ingredient: per tablet: Aspirin (325mg) ...
Drug interaction precautions Do not take this product if you are taking a prescription drug for anticoagulation (thinning the blood), diabetes, gout, or arthristis unless directed by a doctor. ... If you are presently taking a prescription drug, do not take this product without checking with your doctor or other health professionel.
Warnings Children and teenagers should not use this medicine for chicken pox or flu symptoms before a doctor is consulted about Reye Syndrome, a rare but serious illness reported to be associated with aspirin. As with any drug, if you are pregnant or nursing a baby, seek the advice of a health professional before using this product. It is especially important not to use aspirin during the last three months of pregnancy unless specifically directed to do so by a doctor because it may cause problems in the unborn child or complications during delivery. ... Do not take this product if you are allergic to aspirin or if you have asthma, bleeding problems or on a sodium restricted diet. If ringing in the ears or a loss of hearing occurs, consult a doctor before taking any more of this product. If pain persists or gets worse, if new symptoms occur, or if redness or swelling is present, consult a doctor because these could be signs of a serious condition. ...
Alcohol warning If you consume 3 or more alcoholic drinks every day, ask your doctor whether you should take aspirin or other pain relievers / fever reducers. Aspirin may cause stomach bleeding.

Figure 1: Extraits de notices pour les médicaments ANADIN et ALKA-SELTZER

à sa relation possible avec la consommation d'alcool.

Malgré ces premières différences, les notices dans le corpus que nous avons étudié expriment généralement le même type de contenu communicatif. Autrement dit, les *buts communicatives* exprimés par les auteurs de ces notices sont similaires. Toutefois, ce contenu peut être exprimé de nombreuses façons. Une analyse de la variation stylistique dans un corpus de 342 notices pharmaceutiques (Paiva, 2000) montre que deux facteurs importants opposent d'une part *l'abstraction* (ex. utilisation de passifs sans agents) à *l'engagement* (ex. utilisation de pronoms de la 1ère et 2ème personne et de l'impératif), et d'autre part la *référence complète* à la *référence pronominale*. Au delà de cette variation dans l'expression linguistique de surface, nous avons également constaté que des buts communicatifs similaires peuvent être exprimés avec des différences sémantiques plus ou moins importantes. Par exemple, il est admis en médecine qu'on ne doit pas donner d'aspirine à des enfants ou à des adolescents ayant de la fièvre ou d'autres symptômes d'une infection virale (particulièrement la grippe et la varicelle) sans avis médical, car cela peut causer une maladie grave appelée *Syndrome de Reye*. Les avertissements en rapport dans nos deux exemples sont formulés ainsi:

ANA: *Aspirin has also been associated with increased risk of Reye's Syndrome when given to children with a fever. For this reason the use of aspirin is not recommended in children under 12 years of age.*

ALK: *Children and teenagers should not use this medicine for chicken pox or flu symptoms before a doctor is consulted about Reye Syndrome, a rare but serious illness reported to be associated with aspirin.*

Ces deux fragments de texte véhiculent très clairement des différences sémantiques: ALK men-

tionne par exemple qu'un docteur devrait être consulté (*a doctor should be consulted*), alors que ANA indique que l'aspirine n'est pas recommandée pour les enfants âgé de moins de 12 ans (*aspirin is not recommended in children under 12*). De même, ALK fait référence aux enfants et aux adolescents ayant des symptômes de varicelle ou de grippe (*children and teenagers with chickenpox or flu symptoms*), et *ANA* fait référence aux enfants en état de fièvre (*children with a fever*). Ces distinctions sémantiques révèlent des choix différents faits par les auteurs de ces notices. Toutefois, il existe bel et bien un but communicatif commun, qui pourrait être formulé comme suit: *l'aspirine ne doit être administrée que sous supervision médicale aux enfants et adolescents présentant des symptômes d'une infection virale*. Pour des documents de l'importance des notices pharmaceutiques, on peut soutenir le fait que la cohérence dans l'expression des buts communicatifs ainsi que dans la structure du contenu peut aider à une compréhension claire et non-ambigüe. Ceci se retrouve dans des compilations de notices de médicaments ayant été normalisées, comme Le Vidal de la Famille (Vidal, 1998).

Les observations que nous avons faites nous permettent de donner une définition possible pour la normalisation de documents. La normalisation d'un document dans un domaine de discours particulier implique son analyse sous forme de représentation du contenu sémantiquement possible, et la production d'une version normalisée du document à partir de cette représentation. Cette version normalisée exprime un *contenu communicatif prédéfini*, présent sous quelque forme dans le document source, avec une structure et une expression linguistique contrôlées. Le contenu prédéfini révèle des *buts communicatifs*, lesquels devraient idéalement être décrits par un expert du domaine de discours. Un tel but communicatif pour des notices pour des médicaments contenant de l'aspirine mettrait en garde contre le Syndrome de Reye. Produire le texte du document normalisé à partir d'une représentation du contenu permet de générer un message pouvant être considéré comme *l'étalon or* pour l'expression de ce concept (c'est-à-dire une expression claire et largement acceptée adaptée au lecteur). Cela permet également d'obtenir des structures de documents cohérentes ainsi que d'imposer des contraintes terminologiques et stylistiques sur les textes produits.

2.2 Spécification du contenu des documents

Le niveau de spécification du contenu des documents que nous souhaitons utiliser pour la normalisation de documents est donc celui des buts communicatifs (par opposition à un niveau de sémantique exprimable en termes de formes logiques par exemple). Dans le cas de notre exemple précédant concernant le Syndrome de Reye, nous ne souhaiterions n'avoir à spécifier que la présence du but communicatif riskOfReyeSyndrome dans la partie appropriée de la représentation du contenu du document, et donc déléguer la production du texte pour exprimer ce concept à un processus de génération propre.

La représentation du contenu d'un document doit être complète et bien formée. Elle doit être *complète* par rapport à des attentes propres à la *classe de documents* (l'ensemble des documents comparables dans un même domaine de discours). Elle doit être *bien formée* dans le sens où sa structure sémantique doit être possible dans le domaine de discours considéré, et où elle doit respecter les contraintes sémantiques pouvant exister entre ses sous-structures. Par exemple, une telle contrainte pourrait exprimer que le type d'administration pour un médicament dans la sous-structure concernant les modes d'administration doit être compatible avec la forme pharmaceutique du médicament telle que spécifiée dans la sous-structure de description du médicament (de sorte que l'on ne puisse spécifier qu'un médicament sous forme de poudre doive être croqué).

Une telle modélisation de la sémantique des documents ne semble possible que pour des domaines de discours limités mettant en jeu des types de contenu prédéfinis, comme cela semble être le cas pour le corpus que nous avons étudié. Sous ces conditions, spécifier le contenu d'un document peut être vu comme la définition d'un point dans l'espace sémantique du domaine de discours considéré qui identifie de façon unique le document.

2.3 Systèmes de création de documents

On constate un intérêt récent pour la recherche dans le domaine des systèmes de création de documents (Power et Scott, 1998; Brun et al 2000; Hallgren et Ranta, 2000; Coch et Chevreau, 2001). Cette recherche s'est principalement concentrée sur l'obtention des représentations du contenu des documents par interaction avec l'utilisateur du système (*l'auteur*), et la production de versions multilingues du document correspondant à ces représentations spécifiées par les choix de l'auteur. Typiquement, l'auteur doit faire des choix sémantiquement valides dans des *zones actives* qui apparaissent dans le texte en évolution du document, lequel est généré au fur et à mesure de la création du document dans la langue de l'auteur. Les sélections ainsi faites raffinent itérativement la représentation du contenu du document jusqu'à ce que celle-ci soit complète.

Dans le système MDA (Multilingual Document Authoring) (Dymetman et al, 2000) (voir figure 2), la spécification de représentations du contenu des documents bien formées est décrite récursivement dans un formalisme grammatical qui est une variante des Definite Clause Grammars (Pereira et Warren, 1980). Un extrait de grammaire très simple pour une notice pharmaceutique dans le formalisme MDA est donné en figure 3.[3] La première règle se lit ainsi: la structure sémantique leaflet(T,P,D,W,...)[4] est de type patientLeaflet; ce type est constitué d'une structure title, d'une structure presentation, d'une structure directions, d'une structure warnings, et d'une liste de structures n'apparaissant pas dans notre exemple. Les contraintes sémantiques sont établies à l'aide de paramètres de type partagés: par exemple, PersonCategory contraint les structures presentation et warnings (ceci permettant d'exprimer le type de contrainte mentionné plus tôt). La seconde règle indique que riskOfReyeSyndrome et de type warnings(children).[5] Cette règle illustre également la possibilité d'avoir des chaînes de caractères dans les parties droites des règles, ce qui permet d'associer des réalisations textuelles à des représentations du contenu des documents.[6]

3 Un paradigme pour l'analyse de contenu automatique appliquée à la normalisation de documents

Comme nous l'avons vu dans la section précédante, un système comme MDA permet de représenter le contenu de documents bien formés ainsi que de produire des réalisations textuelles à par-

[3] MDA a déjà été utilisé pour la modélisation de notices pharmaceutiques plus complexes (Brun et al 2000).

[4] Les points indiquent des variables qui ne sont pas montrées.

[5] Il s'agit ici d'une simplification pour l'exposition, puisque une telle structure contient habituellement plusieurs avertissements, et qu'elle peut dépendre d'autre paramètres.

[6] Le texte d'un document est obtenu en parcourant sa structure sémantique et en concaténant les chaînes de caractères de façon ascendante, après avoir éventuellement appliqué des contraintes de niveau morphologique qui n'apparaissent pas dans notre exemple (Brun et Dymetman, 2002)

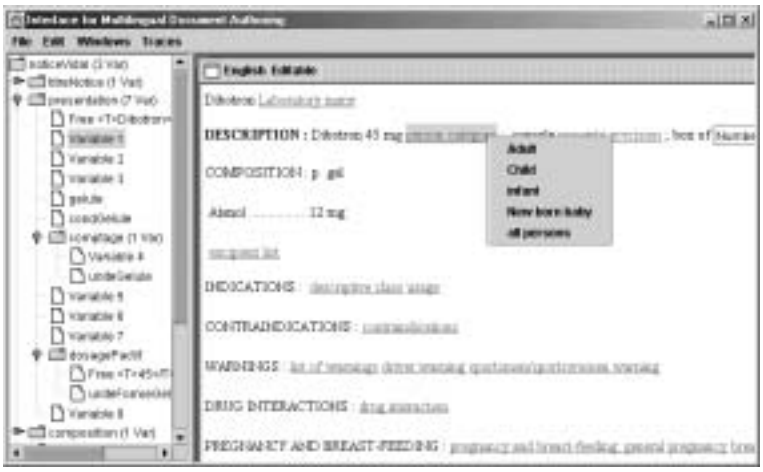

Figure 2: Exemple de création de notice pharmaceutique à l'aide du système MDA

```
leaflet(T,P,D,W,...)::patientLeaflet --->
        T::title,
        P::presentation(PharmaceuticalForm, PersonCategory),
        D::directions(PharmaceuticalForm),
        W::warnings(PersonCategory),
        ... .

riskOfReyeSyndrome::warnings(children) -->
  ['Aspirin should not be given to children and teenagers with a fever or other symptoms of
    a virus infection, especially flu or chickenpox, unless prescribed by a doctor, because
    it may cause a serious illness called Reye's Syndrome.']
```

Figure 3: Extrait de grammaire MDA pour une notice pharmaceutique.

tir de ces représentations. Ainsi, un tel système peut être utilisé pour faire l'énumération de l'ensemble des documents possibles pour la classe de documents modélisée. En rendant la grammaire de génération non-déterministe, on peut obtenir un nombre important de réalisations textuelles associées avec des représentations du contenu des documents, rendant compte en partie de la diversité rencontrée dans les corpus.

Il existe de fortes motivations pour vouloir réutiliser des modèles sémantiques précédemment développés pour de tels systèmes de création de documents pour normaliser des documents existants. La représentation du contenu d'un document à analyser peut être obtenue par la recréation des choix sémantiques, définis dans un modèle sémantique donné, qui sont exprimés dans ce document. Lorsqu'une représentation complète et bien formée est obtenue, le texte de la version normalisée du document peut être produit par la grammaire générative du système de création de documents (éventuellement en plusieurs langues en utilisant des grammaires parallèles).

3.1 Génération inversée floue

L'analyse du contenu des documents est souvent abordée comme un problème d'analyse syntaxique où des représentations sémantiques sont composées à partir de structures syntaxiques (Allen, 1995). Il est toutefois très difficile en pratique de développer des grammaires d'analyse

syntaxique à couverture large qui soient robustes à la variabilité linguistique que l'on peut trouver dans les textes. De plus, faire la correspondance entre une représentation sémantique dérivée d'une structure syntaxique et un but communicatif (notre niveau cible) n'est en général pas trivial. Nous proposons donc un paradigme pour l'analyse du contenu en domaine *limité et bien défini* qui inverse cette vue en se concentrant sur la notion de *génération inversée floue* à partir de structures sémantiques produites par un modèle (Max et Dymetman, 2002).

Un modèle MDA peut être utilisé pour faire l'énumération de l'ensemble des représentations du contenu de documents bien formées pour sa classe de documents, et associer des réalisations textuelles à ces représentations par un processus de génération. Nous appellerons ces réalisations les *documents virtuels* du modèle sémantique, parce que ces documents n'existent pas avant d'avoir été effectivement produits par le modèle. Conceptuellement, pour normaliser un document nous souhaiterions trouver le document virtuel le plus proche du document à analyser en termes des buts communicatifs qu'il exprime. Nous pourrions alors considérer sa représentation du contenu associée comme une approximation de celle du document à analyser. La *génération inversée* intervient à ce niveau, puisque les prédictions faites par la grammaire de génération à partir de représentations bien formées peuvent être utilisées pour mesurer une certaine forme de similarité entre deux documents. Mais puisque la grammaire de génération sous-génère relativement à l'ensemble des textes qui expriment le même contenu communicatif, nous utilisons un méchanisme de *correspondance floue* pour faire l'estimation d'une mesure de similarité qui essaie de rendre compte de la quantité de contenu communicatif partagée par deux textes.

3.2 Similarité de contenu entre deux textes

Nous avons défini notre notion de similarité de contenu à partir du fait, généralement accepté dans le domaine de la recherche d'information, que plus deux textes partagent de termes et de termes liés (ex. *grossesse* et *enceinte*), et plus ils sont succeptibles d'être à propos du même sujet. Le contenu d'un texte peut être grossièrement approximé par un vecteur contenant les formes lemmatisées des termes et leur nombre d'occurrences. Nous appelons un tel vecteur le *profil lexical* d'un texte. Afin de rendre compte de la variation lexico-sémantique, les éléments d'un profil lexical sont en fait des ensembles de synonymes au sens de WordNet (*synsets*), ce qui permet à la fois d'avoir une représentation lexicalement désambiguïsée et d'identifier les termes équivalents (Gonzalo et al, 2000).

La mesure de similarité que nous cherchons à évaluer doit rendre compte de la quantité de contenu communicatif *commun* à deux textes (dans un premier temps, nous ne considérons pas la masse communicative qui est propre à chaque texte[7]). Une telle mesure de similarité peut donc être obtenue par une mesure d'intersection entre les profils lexicaux des deux textes. De plus, les éléments présents dans les profils lexicaux ne participent pas de la même façon à la caractérisation de leur contenu communicationnel, c'est pourquoi la mesure d'intersection est pondérée par une valeur d'*informativité*. Dans notre implémentation initiale, nous donnons une informativité nulle aux mots fonctionnels, et une informativité dérivée de leur fréquence inverse dans un corpus aux ensembles de synonymes (ainsi, un terme (ainsi que ses synonymes) apparaissant peu souvent, comme *fièvre*, partipera davantage à l'intersection qu'un terme apparaissant plus souvent, comme *médicament*). La formule suivante donne cette mesure de similarité. $occs_{L1}(synset)$ est le nombre d'occurrences de $synset$ dans le profil lexical $L1$, et $inf(synset)$

[7]Sous cette hypothèse, le processus de normalisation peut être comparé à un processus d'extraction d'information, ou encore de résumé automatique en domaine contraint.

```
- créer une liste vide de représentations du contenu (OPEN)
- créer une liste vide de candidats (CAND) (représentations du contenu complètes)
- mettre le type 'document' (représentation partielle initiale d'un document) dans OPEN
- répéter jusqu'à ce que N candidats aient été trouvés
   - enlever le premier élément de OPEN
   - si cet élément est une représentation du contenu complète l'ajouter à CAND
   - sinon pour chacun de ses successeurs, calculer leur similarité avec le texte
     source et les insérer dans OPEN par similarité décroissante
```

Figure 4: Algorithme de recherche des structures candidates

son informativité.

$$sim(L1, L2) \quad = \quad \sum_{synset \in L1, L2} min(occs_{L1}(synset), occs_{L2}(synset)) * inf(synset)$$

3.3 Recherche admissible des structures candidates

En pratique, nous ne souhaitons pas comparer l'ensemble des documents virtuels au document à analyser. Nous proposons donc une recherche heuristique qui garantie, grâce à son caractère *admissible* (Nilsson, 1998), que les N premiers documents trouvés sont les N plus similaires avec le texte à analyser. L'hypothèse qui est faite ici est que la mesure de similarité et la valeur de N choisies permettent de garantir que le meilleur candidat pour la normalisation du document à analyser se trouve dans la liste retournée. La procédure est similaire à une analyse syntaxique descendante où des représentations partielles du contenu des documents sont itérativement construites en tenant compte d'une mesure de similarité avec le texte à analyser. La recherche est admissible si elle suit une stratégie en *meilleur d'abord*, et si la fonction d'évaluation utilisée est *optimiste*, c'est-à-dire qu'elle sur-estime la valeur réelle de la similarité entre le document source et *quelque document virtuel pouvant être produit à partir d'une représentation partielle du contenu d'un document donnée*. De plus, la fonction d'évaluation doit décroître au fur et à mesure de la progression de la recherche. Un noeud dans l'espace de recherche est une représentation partielle du contenu d'un document, et ses successeurs peuvent être obtenus en appliquant un pas de dérivation dans la grammaire (une variable non-instanciée dans une représentation partielle d'un document reçoit une valeur compatible avec son type), puis en éliminant les représentations obtenues qui ne respectent pas les contraintes sémantiques imposées par le modèle. L'algorithme donné en figure 4 est une implémentation d'une telle recherche qui examine à chaque itération le noeud avec la meilleure évaluation et retourne la liste des N meilleures structures candidates.

3.4 Similarité de contenu entre une représentation partielle et un texte

La fonction d'évaluation que nous utilisons est l'intersection pondérée entre deux profils lexicaux: l'un est celui du document à analyser, l'autre est celui d'une représentation du contenu partielle. Nous pouvons définir la notion de profil lexical pour les types de la grammaire en propageant les profils lexicaux obtenables pour les terminaux (chaînes de caractères) de la grammaire. Un type donné peut avoir plusieurs réalisations, qui correspondent toutes à une collection de textes virtuels. Le profil lexical d'un type doit donner une mesure du nombre

```
- construire le profil lexical d'un type T
  - pour chacune de ses réalisations REA
    - pour chaque élément dans REA construire son profil lexical s'il n'a pas déjà
      été construit
    - calculer le profil lexical pour REA en sommant les nombres d'occurrences de
      chaque élément
  - calculer le profil lexical pour T en prenant le maximum d'occurrences pour chaque
    élément dans chacune de ses réalisations
  - indiquer le profil lexical de ce type T comme ayant été construit
```

Figure 5: Précompilation des profils lexicaux des types de la grammaire

maximum d'occurrences des membres d'un ensemble de synonymes pouvant être obtenu en dérivant ce type de toutes les façons possibles. Le profil lexical pour la réalisation d'un type (la partie droite d'une règle) peut être obtenu en prenant une union des profils lexicaux de l'ensemble de ses éléments qui somme le nombre d'occurrences pour chaque élément.

Le profil lexical d'un type peut être alors obtenu en prenant pour chaque élément son nombre maximum d'occurrences dans chacune des réalisations pour ce type. Cela reflète le fait que quelle que soit la dérivation qui est faite à partir d'un type, un élément donné ne peut apparaître dans un texte produit à partir de cette dérivation plus d'un certain nombre de fois. Un processus de *précompilation* récursif de la grammaire permet de construire les profils lexicaux pour l'ensemble des types de la grammaires (voir figure 5).[8] Finalement, le profil lexical d'une représentation partielle est l'union (sommant les nombres d'occurrences) des profils lexicaux pour les types de toutes ses variables non-instanciées (les parties non-spécifiées) et de ses fragments de texte (les parties connues). On peut facilement montrer que la mesure de similarité par intersection donnée plus tôt ne peut que décroître ou rester constante au fur et à mesure que des représentations partielles sont raffinées, satisfaisant ainsi les contraintes d'admissibilité.

3.5 Implémentation initiale et critères d'évaluation

Nous avons développé un prototype initial utilisant une grammaire hors-contexte qui est une simplification d'une grammaire MDA pour des notices pharmaceutiques, qui utilise le lemmatizer de notre laboratoire (XRCE, Finite-state linguistic components). Un nouveau prototype supportant le formalisme MDA est en cours de développement dans le cadre de notre travail. Le processus de normalisation peut être inspecté en lisant et comparant les textes retournés avec le texte source. Nous souhaitons mettre en place une procédure d'évaluation automatique qui comparera les représentations du contenu de documents issus d'un corpus de test ayant été 'manuellement' analysés à l'aide de l'interface de MDA avec celles obtenues pour le document source.

[8] La version de l'algorithme donnée est pour une grammaire non-récursive, mais elle peut être adaptée au cas d'une grammaire récursive si l'on autorise une valeur infinie (en pratique, une borne supérieure) pour le nombre d'occurrences d'un mot.

4 Conclusions et perspectives

Nous avons introduit la problématique de la normalisation de documents, et nous avons proposé un paradigme et une implémentation possibles pour l'analyse du contenu de documents utilisant un modèle sémantique et la génération inversée floue. L'approche que nous avons proposée est plus simple à implémenter qu'une grammaire d'analyse syntaxique, mais elle permet néanmoins d'obtenir une certaine robustesse à la variabilité linguistique présente dans les documents analysés. Il nous reste encore à évaluer les limites des techniques simples que nous utilisons pour la mesure du contenu communicatif commun entre textes. Il faut toutefois remarquer que malgré le caractère simple de la mesure de similarité utilisée, beaucoup de fausses analyses ne seront jamais considérées: en effet, seules les représentations du contenu pouvant être produites par le modèle entrent en compétition, ce qui joue déjà pour une part importante dans la discrimination des candidats.

Nous considérons néanmoins l'approche proposée comme une première passe pour la normalisation de documents: une fois qu'un nombre fini de documents virtuels a été isolé, des techniques plus complexes de traitement des langues peuvent être appliquées sur ces documents. Celles-ci peuvent aller par exemple de la simple mesure de proximité entre mots jusqu'à la mise en correspondance de dépendances telles que retournées par un analyseur syntaxique robuste, tout en gardant en compte que l'on cherche avant tout à déterminer la présence de buts communicatifs communs. Finalement, un expert de la classe de documents pourrait intervenir dans une phase de désambiguïsation interactive.

Références

James Allen (1995), *Natural Language Understanding*, 2ème édition, Benjamin/Cummings.

Avi Arampatzis, Th. P. van der Weide, P. van Bommel, C.H.A. Koster (2000), Linguistically Motivated Information Retrieval, *Encyclopedia of Library and Information Science*, Marcel Dekker, Vol. 69, 2000.

Dominique Dupagne, Pauline Groleau, Colette Pecquet, Marie-Catherine Bonjean (1998), Le Vidal de la Famille, OVP Editions du Vidal, Hachette, Paris.

Caroline Brun, Marc Dymetman, Veronika Lux (2000), Document Structure and Multilingual Authoring, Actes de *INLG 2000, Mitzpe Ramon, Israël*.

Caroline Brun, Marc Dymetman (2002), Rédaction Multilingue Assistée dans le Modèle MDA, Dans *Multilinguisme et Traitement de l'Information*, Frédérique Segond éditeur, Hermès.

José Coch, Karine Chevreau (2001), Interactive Multilingual Generation, Actes de *CICLING 2001, Mexico, Mexique*.

Marc Dymetman, Veronika Lux, Aarne Ranta (2000), XML and Multilingual Document Authoring: Convergent Trends, Actes de *COLING 2000, Saarbrucken, Allemagne*.

Julio Gonzalo, Felisa Verdejo, Irina Chugur, Juan Cigarrán (1998), Indexing with WordNet Synsets Can Improve Text Retrieval, Actes du *COLING/ACL-98 Workshop on the Usage of WordNet in Natural Language Processing Systems*, Montréal, Canada.

T. Hallgren et A. Ranta (2000), An Extensible Proof Text Editor, M. Parigot et A. Voronkov éditeurs, *Logic for Programming and Automated Reasoning*, LPAR'2000, Springer Verlag, Heidelberg.

Aurélien Max et Marc Dymetman (2002), Document Content Analysis through Inverted Generation, *AAAI Spring Symposium on Using (and Acquiring) Linguistic (and World) Knowledge for Information Access, Stanford University, Etats-Unis*.

Nils J. Nilsson (1998), *Artificial Intelligence: a New Synthesis*, Morgan Kaufmann.

Daniel S. Paiva (2000), Investigating Style in a Corpus of Pharmaceutical Leaflets: Results of a Factor Analysis, Actes du *ACL 2000 Student Research Workshop, Hong Kong*.

Fernando Pereira, David Warren (1980), Definite Clauses for Language Analysis, *Artifical Intelligence*, Vol. 13, 1980.

Richard Power, Donia Scott (1998), Multilingual Authoring using Feedback Texts, Actes de *COLING/ACL-98, Montréal, Canada*.

Xerox Research Centre Europe, Finite-state linguistic components,
http://www.xrce.xerox.com/research/mltt/fsnlp.