TALN 2004, Fès, 19–21 avril 2004

Extraction d'information en domaine restreint pour la génération multilingue de résumés ciblés

Caroline Brun, Caroline Hagège
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan France
Caroline.Brun@xrce.xerox.com Caroline.Hagege@xrce.xerox.com

Résumé - Abstract

Dans cet article nous présentons une application de génération de résumés multilingues ciblés à partir de textes d'un domaine restreint. Ces résumés sont dits ciblés car ils sont produits d'après les spécifications d'un utilisateur qui doit décider *a priori* du type de l'information qu'il souhaite voir apparaître dans le résumé final. Pour mener à bien cette tâche, nous effectuons dans un premier temps l'extraction de l'information spécifiée par l'utilisateur. Cette information constitue l'entrée d'un système de génération multilingue qui produira des résumés normalisés en trois langues (anglais, français et espagnol) à partir d'un texte en anglais.

We present an application of oriented multilingual summarization from domain specific texts. These summaries are oriented because the user has to define in a first step the kind of information he/she wants to be present in the final summary. In order to acheive this task, a first step of information extraction is performed. This extracted information which corresponds to the user's specification is then the input of a multilingual generator that produces the desired summaries in three languages (english, french and spanish) from an english input text.

Mots-clefs – Keywords

Résumé multilingue ciblé, extraction d'information, génération multilingue.
Multilingual oriented summaries, information extraction, multilingual generation.

Introduction

Dans cet article nous présentons une méthode et un prototype qui montrent comment extraire de l'information à partir de textes tout venant appartenant au domaine de la chimie et comment générer des résumés multilingues. Les résumés que nous obtenons sont dits ciblés dans la mesure où ils ne s'attachent qu'à exprimer de l'information que l'utilisateur aura préalablement définie. Il ne s'agit donc pas ici de résumé au sens habituel du terme puisqu'un texte qui ne contiendra pas l'information que nous souhaitons extraire ne donnera aucune sortie alors qu'un texte contenant exclusivement de l'information que nous voulons extraire produira en sortie un autre texte dans lequel l'intégralité de cette information sera reprise et reformulée. L'expérience que nous avons menée s'appuie sur un ensemble de textes décrivant des produits toxiques (voir section 1.1). Dans une première étape, les documents sont analysés, produisant en sortie une représentation normalisée de l'information considérée comme pertinente par l'utilisateur. Dans une deuxième étape, cette information normalisée sert de base à un système de génération qui produit en sortie un ensemble de textes en trois langues (français, espagnol et anglais) générés en parallèle et exprimant en langue naturelle l'information normalisée extraite du texte initial. Après une première partie présentant le corpus sur lequel nous avons travaillé, nous expliquons la manière dont nous extrayons l'information normalisée à partir des textes. C'est sur ce point essentiel que l'approche que nous présentons ici se distingue de celle proposée par (Lenci et al. 2002) qui produit des résumés multilingues à partir de requêtes d'un utilisateur faites sur un texte. En effet, la sélection des phrases pertinentes dans l'approche de (Lenci et al. 2002) n'inclue pas l'étape de normalisation lexicale et syntaxique que nous présenterons plus avant et qui nous permettra de normaliser en une représentation similaire des expressions linguistiques véhiculant un sens proche, même si ces expressions ont une réalisation très différente dans les textes. La deuxième partie décrit comment l'information normalisée extraite en première partie est utilisée pour générer les résumés multilingues. Enfin, dans la mesure où les textes résumés produits sont écrits en langue naturelle, on peut les comparer avec le texte initial, ce qui permet d'évaluer de manière plus simple l'extraction ciblée d'information.

1 Extraction d'information ciblée

1.1 Caratéristiques du corpus

Le corpus sur lequel nous travaillons fait partie du domaine de la chimie pour un usage grand-public. Il est composé d'un ensemble de textes en anglais provenant de l'ATSDR (*Agency of Toxic Substance and Disease Registry*). Ces textes sont des paragraphes contenant en moyenne six à sept phrases d'une collection qui présente des produits toxiques, leur provenance, leurs caractéristiques, leur utilisation, leurs effets, etc. L'information véhiculée par ces textes est relativement homogène et ces textes présentent de nombreuses reformulations et paraphrases.

Par exemple, dans les textes concernant respectivement *l'acétone* et *l'acroléïne* les propriétés physiques de ces produits, qui sont pratiquement identiques, sont cependant exprimées de manière très différente :

It evaporates easily, is flammable, and dissolves in water. (texte sur *l'acétone*)
It dissolves in water very easily and quickly, changes to a vapor when heated. It also burns easily. (texte sur *l'acroléïne*)

De plus nous pouvons trouver dans le même texte des redondances, par exemple dans le texte décrivant le *2-Butanone* :
It is also present in the environment from natural sources.
et plus loin :
2-Butanone occurs as a natural product.

L'un des enjeux de notre normalisation est d'obtenir une représentation uniforme à partir de ces expressions linguistiques véhiculant un sens très proche.

1.2 Normalisation d'information ciblée

Dans un premier temps nous avons sélectionné l'information que nous souhaitons extraire à partir de ces textes. Cette information concerne l'aspect, les propriétés physiques, les synonymes, l'utilisation et l'origine de ces produits. Notre but est de produire une représentation normalisée de toute l'information concernant ces thèmes et véhiculée dans les textes par des expressions linguistiques variées.

Les prédicats ci-dessous sont utilisés pour la représentation normalisée :

- **DESCRIPTION_NATURE/2**. Ce prédicat est instancié lorsque dans le texte la nature du produit est spécifiée. Par exemple, **DESCRIPTION_NATURE(ammonia,gas)** est obtenu à partir de toute expression linguistique véhiculant l'information que le produit chimique *ammoniaque* est un gaz.

- **DESCRIPTION_COLOUR/2**. Ce prédicat est instancié par des expressions linguistiques indiquant la couleur du produit. Par exemple, **DESCRIPTION_COLOUR(antimony,silvery-white)** indique que *l'antimoine* est un produit blanc-argenté.

- **DESCRIPTION_SMELL/2**. Ce prédicat est instancié lorsque dans le texte l'odeur du produit chimique est indiquée, i.e. **DESCRIPTION_SMELL(1,3-butadiene,gasoline-like).**

- **SYNONYM/2**. Ce prédicat à deux arguments lorsqu'il est instancié donne le synonyme du produit décrit par l'article, i.e. **SYNONYM(acetone,dimethyl ketone).**

- **PROPERTY/5**. Le prédicat **PROPERTY** est le résultat de la normalisation d'expressions linguistiques qui décrivent une propriété physique ou chimique du produit. Par exemple **PROPERTY(acrolein,dissolve,water,in,NONE)**[1] nous indique que *l'acroleïne* est soluble dans l'eau (instantiation des quatre premiers arguments du prédicat). Le dernier argument du prédicat est représenté ici par une variable car dans le texte initial la manière dont se produit la dissolution n'est pas spécifiée dans le texte. Dans d'autres cas, il sera possible de trouver des variables sur les arguments portant l'information sur la localisation de la propriété. Par exemple, **PROPERTY(acrolein,burn,NONE,NONE,easily)** est obtenu après le traitement du texte sur *l'acroleïne* dans lequel il est seulement indiqué que ce produit est facilement inflammable.

- **ORIGIN/5**. Ce prédicat instancié contient de l'information sur l'origine du produit. Par exemple, **ORIGIN(ammonia,manufactured,NONE,NONE,NONE)** exprime que *l'ammoniaque* est un produit fabriqué par l'homme et **ORIGIN(amonnia,natural,soil,in,bacteria)** exprime que

[1] La préposition est conservée dans ce prédicat car elle est utilisée dans la phase de génération.

ce même produit est également un produit naturel produit par des bactéries et que l'on peut trouver dans le sol.

- **USE/6** est le résultat de la normalisation de texte exprimant l'utilisation du produit chimique décrit. Par exemple, **USE(benzidine,NONE,NONE,produce,dye,before)** exprime que la *benzidine* était utilisée dans le passé (dernier argument) pour produire de la teinture (4eme et 5eme arguments) et **USE(ammonia,smelling_salts,in,NONE,NONE,now)** indique que *l'ammoniaque* est utilisée actuellement (dernier argument) dans des sels (2eme et 3eme arguments) dans un but non spécifié (arguments variables).

A tout ces prédicats peut être ajouté le suffixe **_NEG** lorsque les expressions linguistiques ayant permis leur instanciation sont niées.

1.3 Mise en oeuvre

Le système d'extraction d'information ciblée prend en entrée un texte décrivant un produit toxique et donne en sortie un ensemble de prédicats totalement ou partiellement instanciés. Le traitement s'effectue grâce à un seul outil, XIP (Aït-Mokhtar et al. 2002) qui prend en charge toutes les étapes du traitement, de la segmentation du texte à la production des prédicats. Ce traitement s'effectue selon deux axes décrits ci-dessous.

1.3.1 Analyse morpho-syntaxique normalisée robuste

Cette phase d'analyse consiste en une analyse grammaticale en dépendance, analyse qui est normalisée en utilisant des propriétés de la morphologie dérivationnelle (équivalence entre noms et verbes de même famille morphologique), des propriétés de structures syntaxiques (l'équivalence passive-active, alternances verbales), de l'analyse syntaxique profonde (récupération des sujet des infinitives). Pour plus de détail sur l'analyse syntaxique normalisée robuste voir (Hagège, Roux, 2003). A la sortie de cette étape, une expression comme *Antimony is mixed with other metals* est représentée comme suit :
SUBJ-N(mix,SOMEONE)
OBJ-N(mix,Antimony)
OBJ-N(mix,metal)
Cette représentation signifie que l'action est portée par le lemme *mix*, que le sujet normalisé de ce verbe est non spécifié et que ce verbe a deux objets normalisés représentés respectivement par les lemmes *Antimony* et *metal*. Il faut également noter que pour cette analyse, l'unité de traitement fournie à l'analyseur n'est pas la phrase, mais un paragraphe complet décrivant le même produit.

1.3.2 Traitement orienté par l'application et le domaine

A l'étape d'analyse normalisée et générale mentionnée au point précédent vient se greffer une analyse dépendante du domaine et de l'application. Ce traitement est imbriqué dans le précédent à plusieurs niveaux du traitement.

Régles de segmentation spécifiques: Les chaînes de caractères désignant les entités chimiques peuvent présenter des caractères qui sont habituellement considérés comme des séparateurs

dans la langue standard (i.e. *2,3-Benzofuran*). Afin de pallier ce problème nous avons élaboré des règles locales spécifiques qui s'appliquent sur le texte segmenté et analysé morphologiquement.

Règles de désambiguisation spécifiques: Dans la mesure où nous travaillons dans un domaine particulier, certaines des ambiguïtés pouvant apparaître dans la langue générale n'existent plus dans ce contexte. Par exemple, le mot *sharp* peut être a priori un nom (dièse) ou un adjectif (piquant, pointu). Or dans le contexte de la chimie, nous pouvons rejeter l'étiquette nom pour cette suite de caractères et ne considérer que l'étiquette adjectif. Ces règles spécifiques viennent s'appliquer juste avant l'application des règles d'étiquettage pour la grammaire générale.

Ajout d'information de sémantique lexicale: Pour les besoins de l'application, nous avons typé à l'aide de traits des éléments du lexique tels que les éléments chimiques, les noms de couleur les plus courants, etc.

Meilleur traitement de la coordination: Le corpus contient de longues chaînes d'éléments coordonnés et particulièrement des coordinations dans lesquelles le dernier élément coordonné est précédé par une virgule et une conjonction de coordination. Connaissant cette caractéristique et tirant parti de l'information de sémantique lexicale supplémentaire rajoutée sur certaines entrées, la coordination est traitée avec d'avantage de précision.

Résolution d'anaphore ad-hoc: Le corpus sur lequel nous travaillons présente une particularité intéressante qui est celle de ne poser aucun problème pour la résolution des anaphores pronominales. On peut, avec une marge d'erreur extrêmement faible considérer que tous les pronoms *it* et tous les possessifs réfèrent à l'entité chimique décrite dans le texte.

Relations lexico-structurales adaptées au domaine: Afin de pouvoir obtenir une même représentation pour des expressions linguistiques équivalentes nous avons établi des relations lexico-structurales entre des éléments du lexique puis nous avons crée des règles qui exploitent ces relations nous permettent d'obtenir une représentation normalisée.

Ainsi, nous voulons pouvoir relier l'adjectif *flammable* et le verbe *burn* et dire que si l'adjectif modifie un nom d'entité chimique, alors la propriété physique de cet entité est portée par le verbe *burn*. Une relation de même type que celle existant entre *flammable* et *burn* existe entre *soluble* et *dissolve*, *volatile* et *evaporate* etc.

Nous avons crée de manière déclarative des relations entre ces paires de mots, et ces relations pourront être exploitée par notre système au même titre que des relations de dépendances qui auront été calculées lors du traitement.

Les relations que nous avons créees sont les suivantes:

- **ISAJ** lie un adjectif et un verbe lorsque le verbe peut être paraphrasé par *BE+adjective* (ex. **ISAJ(soluble,dissolve)**).
- **TURNTO** lie un nom et un verbe lorsque le verbe peut être paraphrasé par *TURN TO+noun* (ex. **TURNTO(liquid,liquefy)**).
- **HASN** lie un nom et un verbe lorsque le verbe peut être paraphrasé par *HAVE+noun* (ex. **TURNTO(dissolve,solubility)**).
- **SYNO** lie deux mots de même catégorie morpho-syntaxique quand le premier est syn-

onyme du second[2].

règles de normalisation: Les relations mentionnées ci-dessus sont ensuite exploitées par des règles XIP. Ces règles exploitent à la fois des relations de dépendances syntaxiques normalisées et des relations de mise en correspondance permettant ainsi de déduire de nouvelles relations.

Ainsi, on peut déduire à partir de l'expression linguistique *aniline is soluble* le prédicat **PROPERTY(aniline,dissolve,NONE,NONE,NONE)** sachant que :
une relation syntaxique **ATTRIB(aniline,soluble)** a été calculée par l'analyseur,
il existe une relation lexicale **ISAJ(soluble,dissolve)**,
le lemme *dissolve* porte un trait indiquant qu'il s'agit d'une verbe exprimant une propriété physique.

En conclusion de cette section, nous avons montré comment en étendant un analyseur syntaxique normalisé, nous sommes capable de normaliser de l'information ciblée à partir de textes. La section suivante montre comment à partir de cette information normalisée l'on peut générer des textes exprimant dans trois langues différentes le contenu informatif extrait.

2 Génération

Comme évoqué dans l'introduction de l'article, les résultats de la phase d'analyse, c'est-à-dire une liste de prédicats "sémantiques", sont utilisés comme source d'information d'un générateur qui délivre des versions normalisées des textes analysés.

A cette fin, nous utilisons un système de rédaction assistée de document, MDA (Multilingual Document Authoring), pour lequel une grammaire spécifique est développée. MDA est un système de génération interactif: l'utilisateur de ce système spécifie interactivement le contenu du document jusqu'a sa complétion.

Cependant, dans le contexte du travail présenté dans cet article, MDA est utilisé comme un système de génération automatique(Reiter, Dale, 2000), car l'information nécessaire à la production des documents est automatiquement dérivée à partir des corpus analysés. Ainsi, l'intervention de l'utilisateur n'est plus nécessaire.

Aprés une brève présentation du système, nous montrons comment ce dernier est utilisé pour produire des versions normalisées des textes du corpus initial.

2.1 Présentation du systéme MDA

MDA est un système interactif pour l'aide à la rédaction de documents contrôlés multilingues ou monolingues. La structure de ces documents ainsi que leur cohérence sémantique sont sous le contrôle du systême. Lorsqu'un seul choix ne mène pas à un document valide, le système met automatiquement à jour le document en générant le texte approprié : dans un système comme MDA, la génération automatique est un effet de bord de la rédaction assistée.

Cet outil étend la syntaxe conventionnelle d'éditeurs SGML ou XML, car les choix sémantiques

[2]Dans la mesure où nous travaillons dans un domaine restreint, l'ambiguité sémantique ne pose ici pas de problème.

peuvent s'appliquer jusqu'au niveau des unités lexicales. De plus, le système permet de spécifier des *dépendances* entre des parties distantes du document : ainsi, une modification dans une partie donnée du document peut-être reflétée dans une autre partie, physiquement éloignée (dépendances à longue distance).

Le contenu du document est décrit par un formalisme grammatical appelé "grammaires d'interaction" (GI), qui est dérivé des grammaires à clauses définies de Prolog. Nous renvoyons à (Brun, Dymetman, 2002), qui donne une description détaillée du système et de ses applications. En particulier la syntaxe des règles GI est définie dans cet article.

2.2 Génération automatique de textes normalisés

Afin de générer les textes cibles, nous avons développé une grammaire GI qui couvre la structure et le contenu de cette classe de document. Le développement est réalisé en deux étapes :

- développement de la grammaire de "réalisation": ce premier ensemble de règles permet la réalisation linguistique des différents concepts couverts par le corpus. Ces règles visent à produire un court paragraphe dont le but de communication est la description d'une substance toxique (en particulier de quelle substance il s'agit, quelles sont ses caractéristiques et propriétés physiques, quelle est son origine, quels sont ses synonymes, pourquoi est-elle utilisée).

  Comme mentionné dans la section 1.1, l'étude du corpus a mis en évidence des ensembles de paraphrases qui véhiculent le même sens. Nous avons donc choisi comme unités linguistiques à générer ce qui nous semblait les réalisations les plus appropriées d'un ensemble de paraphrases d'un même concept. Nous avons sélectionné les expressions les plus fréquentes apparaissant dans le corpus de développement. Par exemple, une phrase comme *It is flammable* présente dans le corpus sera normalisée en *It burns easily* dans le texte généré.

- développement de la grammaire spécifique au domaine : le second ensemble de règles encode la connaissance liée au domaine de l'application, dans notre cas il s'agit des différentes caractéristiques de la substance toxique. Cet ensemble de règles peut se comparer à une base de connaissance (Dymetman2002) et même s'il est limité par le nombre de substances décrites dans le corpus, il est a priori illimité.

Lors des expériences antérieures réalisées dans le cadre de MDA, le développement de règles était complètement manuel. Dans le cas de cette expérience, nous utilisons une méthode automatique d'extraction et de normalisation de la connaissance : les résultats de ce processus sont automatiquement convertis en un ensemble de règles de grammaires qui décrivent la connaissance liée au domaine.

De plus, ces grammaires ont la particularité de n'associer qu'un seul document à une substance toxique donnée : l'utilisateur n'interagit plus avec le système de rédaction assistée, car les règles spécifiques au domaine couvrent l'information requise pour permettre la génération des descriptions. Ceci implique donc l'utilisation de MDA comme un générateur de texte. Prenons l'exemple du prédicat USE (utilisation de la substance):
USE(SUBSTANCE,PROD1,PREP,USEV,PROD2,TIME). Dans le cas de *l'acétone*, l'analyse fournit la liste suivantes de prédicats:

USE(acetone,NONE,NONE,make,plastic,now)
USE(acetone,NONE,NONE,make,fiber,now)
USE(acetone,NONE,NONE,make,drug,now)
USE(acetone,NONE,NONE,make,other chemical,now)
USE(acetone,NONE,NONE,dissolve,other substance,now)

A partir de cette liste, un ensemble de règles GI contraignant les usages possibles de *l'acétone* est construit. De plus, nous créons un ensemble de règles lexicales GI qui encodent le fait que *make, dissolve* sont des verbes exprimant une utilisation, que *drug, fiber, other chemical, plastic, other substance*, sont des produits, etc. Nous procédons de la même manière pour tous les types de prédicats extraits par l'analyseur syntaxique (COLOUR(SUBSTANCE,COLOUR), SMELL(SUBSTANCE,ODOR), etc).

Lorsqu'elle est combinée avec la grammaire de réalisation, cette information nous permet de générer le texte normalisé correspondant au texte initial.

Le système MDA étant initialement conçu pour la rédaction assistée de documents multilingues, nous avons exploité cette capacité et créé des grammaires parallèles à celle présentée précédemment, pour le français et l'espagnol. Ces grammaires partagent la même structure sémantique et permettent des réalisations linguistiques propres à chaque langue.

La figure 1 montre la sortie de la phase d'analyse sous forme de prédicats sémantiques, tandis que la figure 2 montre les textes normalisés générés en français, anglais et espagnol, pour le même produit toxique, l'acétone.

Figure 1: Résultat de la phase d'analyse pour l'acétone

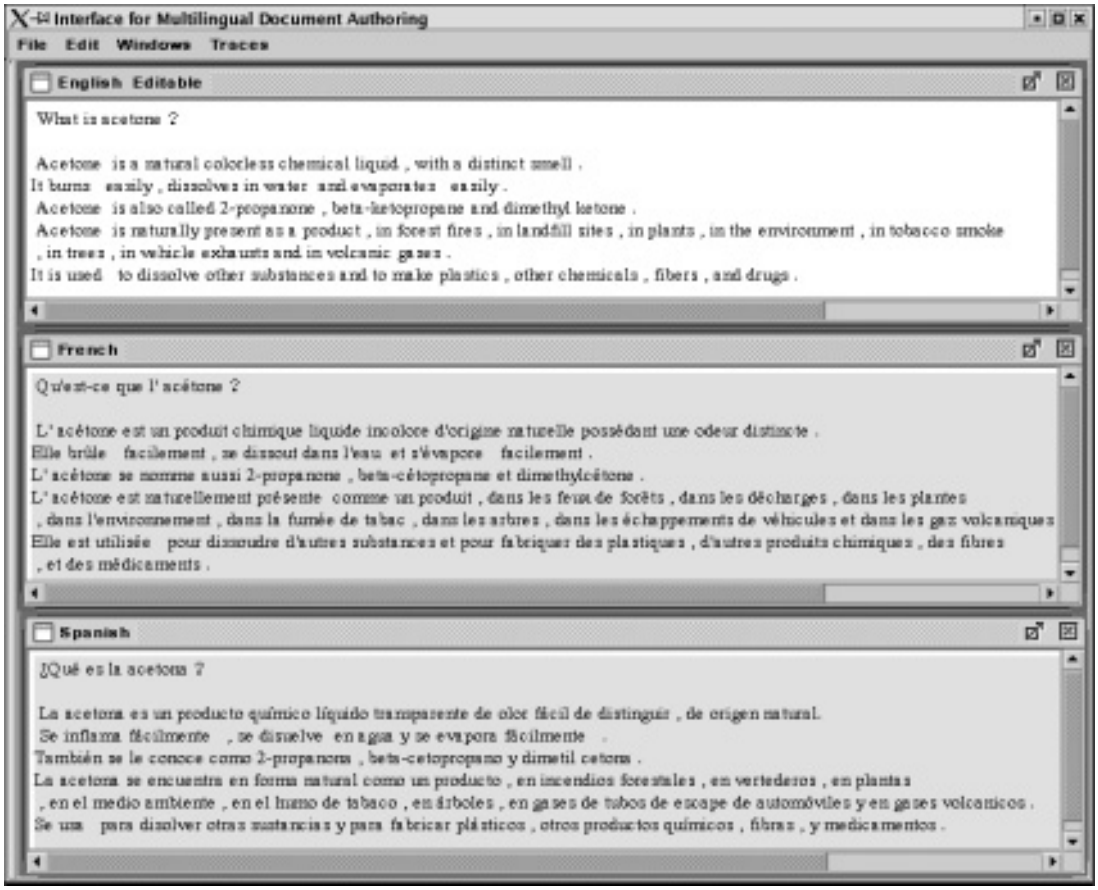

Figure 2: Génération multilingue des textes normalisés

3 Evaluation et Conclusion

L'évaluation de ce travail n'est pas une tâche facile dans la mesure où plusieurs composantes rentrent en jeu. Nous avons ici choisi d'évaluer dans un premier temps les performances de la première partie du traitement (Extraction d'information ciblée). Cette évaluation effectuée sur une collection de 30 textes du domaine qui n'ont pas été préalablement traités par le système a donné les résultats suivants en terme de précision et de rappel. De part l'approche que

Précision	Rappel	FScore
.96	.65	.77

nous avons adoptée toute l'information extraite apparaît nécessairement dans les résumés finaux. Par ailleurs, nous adoptons les critères définis par (Hartley et al. 2001) pour l'évaluation d'une application de génération multilingue et nous pouvons faire les constatations suivantes: **L'acceptabilité** des résumés produites est optimale dans la mesure où la combinatoire des expressions linguistiques possibles est contrôlée et prévue à priori dans la grammaire MDA. Pour ces mêmes raisons le critère de **grammaticalité** des résumés obtenus est également optimale. Enfin, la notion de **couverture** est en fait un reflet des résultats de la phrase d'extraction d'information, puisque tout prédicat extrait dans la première phase aura une réalisation en langue naturelle.

Le travail décrit ici nous présente une méthode et un prototype effectif capable de produire

des résumés ciblés d'une très grande précision à partir de textes appartenant à un domaine restreint. Nous envisageons de porter cette méthode à d'autres domaines et de réfléchir comment l'information linguistique nécessaire pour mener à bien cette tâche peut être acquise de manière partiellement automatique afin de minimiser le temps de développement.

Références

Aït-Mokhtar S., Chanod J-P. and Roux C. (2002), Robustness beyond shallowness: incremental dependency parsing, *Special Issue on Robust Methods in Analysis of Natural Language Data*, Vol. 8, 121-144.

Brun C., Hagège C. (2003), Normalization and Paraphrasing Using Symbolic Methods, Proceeding of the *Second International Workshop on Paraphrasing. ACL 2003, Sapporo, Japan*

Brun C., Dymetman M. (2002), Multilinguisme et traitement de l'information *Rédaction multilingue assistée dans le modèle MDA, pages 129–152 - Traité des sciences et techniques de l'information. Hermès.*

Cancedda N. (1999), Text Generation from MUC Templates. In *Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99)*

Marc Dymetman. 2002. Text authoring, knowledge acquisition and description logics. In *Proceedings of Coling 2002*, Taiwan.

Hartley A., Scott D., Bateman J., Dochev D. (2001), AGILE - A system for Multilingual Generation of Technical Instructions.

Kossein L., Beauregard S., Lapalme L. (2001), Using Information Extraction and Natural language Generation to Answer E-mail. In *Proceedings of the 5th International Conference on Application of Natural Language to Information Systems. Versailles, France.*

Brun C., Dymetman M., Lux V. (2000), Document structure and multilingual authoring. In *Proceedings of the First International Natural Language Generation Conference (INLG'2000), Mitzpe Ramon, Israel* 24-31.

Hagège C., Roux C. (2003), Entre syntaxe et sémantique: Normalisation de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes. Actes de *TALN 2003, Batz-sur-Mer, France*

Lenci, A. Bartolini, R., Calzolari N., Agua A., Buseman S., Cartier E., Chevreau K., Coch J. (2002), Multilingual Summarization by Integrating Linguistic Resources in the MLIS-MUSI Project. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02), May 29-31, Las Palmas, Canary Islands, Spain, 2002*

Reiter E., Dale R. (2000), Building Natural Language Generation Systems. *Studies in Natural Language Processing, Cambridge University Press.*