TOTh 09

Terminologie & Ontologie : Théories et Applications

Actes de la troisième conférence TOTh - Annecy - 4 & 5 juin 2009

Institut Porphyre
Savoir et Connaissance

Dans un monde où la communication et le partage d'information sont au cœur de nos activités, les besoins en terminologie se font de plus en plus pressants. Il est devenu impératif d'identifier les termes employés et de les définir de façon consensuelle et cohérente tout en préservant la diversité langagière.

La terminologie, en tant que discipline scientifique, se fonde sur une conceptualisation d'un domaine et sur les mots pour en parler. Elle se doit donc de concilier un point de vue linguistique et un point de vue ontologique. Elle doit également, dans une société numérique où les connaissances constituent la principale richesse, pouvoir être opérationnalisée à des fins de traitement de l'information.

Les conférences TOTh se situent dans le prolongement des colloques annuels de la Société française de terminologie organisés en décembre à Paris (Ecole normale supérieure de la rue d'Ulm). Planifiées à mi-parcours, au mois de juin à Annecy (Polytech'Savoie), elles en complètent l'offre et proposent des conférences avec appel à communications, comité de lecture et publication des actes.

Les conférences TOTh ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent à la fois de la terminologie et de l'ontologie et, de façon plus générale, de la langue et de l'ingénierie des connaissances. Elles se veulent un lieu d'échange et de partage où sont exposés problèmes, solutions et retours d'expériences tant sur le plan théorique qu'applicatif ; ainsi que les nouvelles tendances et perspectives des disciplines associées : terminologie, langues de spécialité, linguistique, intelligence artificielle, systèmes d'information, ingénierie collaborative, etc.

Christophe Roche, Président du Comité Scientifique

http://www.porphyre.org

Institut Porphyre
Savoir et Connaissance

ISBN 978-2-9536168-0-4
EAN 9782953616804

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2009. *Actes de la troisième conférence TOTh - Annecy – 4 & 5 juin 2009*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2009
ISBN 978-2-9536168-0-4
EAN 9782953616804
© Institut Porphyre, *Savoir et Connaissance*

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2009

Annecy – 4 & 5 juin 2009

avec le soutien de :

- Société française de terminologie
- Association Européene de Terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

http://www.porphyre.org

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Pierre Blanc	EDF SEPTEN
Danièle Bourcier	CNRS, CERSA Paris
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille III
Viviane Cohen	France Télécom, Paris
Rute Costa	Professeur, Université Nouvelle de Lisbonne
Luc Damas	MCF, Université de Savoie
Sylvie Després	MCF, Université Paris XIII
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section de terminologie
Jean-Yves Gresser	ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Jean-Paul Haton	Professeur, Université de Nancy 1
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Université Paris XIII
Widad Mustafa	Professeur, Université de Lille III
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université du Québec en Outaouais
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Dès la troisième édition, les conférences TOTh ont trouvé une structuration qui traduit bien à la fois le caractère scientifique et pluridisciplinaire de la terminologie et l'intérêt de notre communauté pour d'autres domaines partageant des préoccupations communes.

Ainsi, la conférence d'ouverture a été donnée par une personnalité invitée issue d'une discipline différente de la nôtre – ici la phylogénèse – mais pour laquelle le langage et la pensée jouent également un rôle primordial.

Les contributions se sont réparties naturellement, et par le jeu des évaluations de façon équitable, en trois groupes ayant donné lieu à trois sessions.

Le premier groupe a rassemblé les articles portant principalement sur la dimension linguistique de la terminologie. Ont été abordés l'extraction terminologique à partir de dictionnaire, la place accordée aux corpus dans la construction de terminologies, l'acquisition de connaissances à partir de textes et l'apport des ressources linguistiques issues du web.

La deuxième session s'est donc logiquement intéressée à la dimension conceptuelle de la terminologie. Les notions de concept, de relation, d'ontologie ont été au cœur des présentations portant sur les cartes conceptuelles pour les bibliothèques numériques, les relations dynamiques et les graphes conceptuels, l'alignement d'ontologies et l'accès multilingue aux ontologies.

Enfin, la troisième session a été consacrée à la présentation de plusieurs applications terminologiques pour des secteurs aussi différents que l'ingénierie nucléaire, l'informatique, le domaine bancaire ou l'agriculture biologique. Il est à souligner que ces applications ont permis d'aborder différents points théoriques tels que la variation terminologique, la diachronie ou la structure des dictionnaires.

La richesse des débats qui ont animé ces deux jours de conférence – chaque présentation, questions comprises, s'est vue allouer plus de quarante cinq minutes de temps de parole – a été certainement une des plus belles récompenses pour les participants de TOTh 2009.

Christophe Roche

Président du Comité Scientifique

Table des matières

SESSION 3

Extraction des connaissances orientées évolution dans les textes de brevets

Kata Gábor, François Rousselot, François de Bertrand de Beuvron, Denis Cavallucci, Dildar Wu

Résumé : Nous présentons une nouvelle approche pour l'extraction de connaissances à partir des brevets à l'usage des ingénieurs experts d'un domaine. Celle-ci est basée sur la Conception Inventive (CI dans la suite). Une ontologie de la CI aide à préciser la tâche et à définir les informations à extraire. Les résultats obtenus peuvent être directement utilisés par un concepteur, en lui donnant une vision élargie des inventions déposées dans son domaine. Ils seront également exploités comme données d'entrée d'un logiciel développé au sein du LGECO, nommé TrizAcquisition qui permet de mener une étude de CI détaillée sur un artefact donné.

Mots-clés : Extraction d'information, extraction de connaissances, brevet, invention, Conception Inventive, TRIZ

1. Introduction

1.1. Analyse des brevets

Le travail des ingénieurs en Conception Inventive lors de la conception d'une invention consiste souvent à améliorer un artefact existant. Pour ce faire, les ingénieurs ont besoin de connaître l'état de l'art, et en particulier les caractéristiques des produits proches qui existent dans le domaine. Ils ont donc besoin d'accéder à des bases de données de brevets. Celles-ci inventorient la quasi-totalité des brevets existants. Elles disposent de fonctions de recherche limitées basées principalement sur des mots clés. Il est urgent de développer des outils de Traitement Automatique des Langues pour permettre de lancer des requêtes plus fines et obtenir des résultats pertinents et sémantiquement structurés.

1.2. Le modèle de connaissances : la Conception Inventive

Nous travaillons avec des experts praticiens de la "Conception Inventive", issue de la TRIZ[1]. La TRIZ, créée dans les années 50 en ex-URSS et affinée jusqu'à la disparition d'Altshuller (Altshuller 1999) en 1988, voit le processus d'invention comme la résolution d'une contradiction. Nous décrivons ci-dessous les éléments de cette théorie pertinents pour l'extraction de connaissances, et renvoyons le lecteur à (Zanni *et al.* 2009) pour une présentation plus générale.

Nous avons élaboré et formalisé un modèle des connaissances de la Conception Inventive qui rend compte de notre point de vue sur cette théorie et qui en propose une extension. La TRIZ est une théorie de l'évolution des artefacts, elle part du postulat que tout objet conçu par l'homme est le résultat d'une évolution guidée par des lois objectives et la création d'une invention comme résultant d'une impossibilité pour l'objet d'évoluer en cohérence avec ces lois. Une de ses caractéristiques qu'on veut faire évoluer est bloquée par un conflit d'origine technique ou physique. La théorie propose alors une méthode pour formuler précisément ce conflit en termes d'une contradiction qui peut s'énoncer de la manière suivante :

Soit trois paramètres P1, P2, et P3 de l'objet à modéliser. P1 peut prendre les valeurs opposées A ou . or, si P1 prend la valeur A, P2 est améliorée,

1 TRIZ propose une démarche très dirigée pour concevoir de façon inventive en exploitant les analogies entre solutions issues de domaines différents et induit une vue très spécifique sur le processus de l'invention.

mais P3 est dégradée, inversement, si P1 prend la valeur , P2 est dégradée, alors que P3 est améliorée.[2]

Dans une application, une fois qu'un ensemble réduit de contradictions importantes est clairement identifié, TRIZ fournit à l'inventeur des techniques et des bases de connaissances qui lui permettent de générer des Concepts de Solutions. La solution sera qualifiée d'*inventive* si elle est nouvelle et ne constitue pas un compromis[3].

La TRIZ dans sa version actuelle ne concerne que des problèmes ne comportant que peu de contradictions. Nous avons étendu la théorie à des problèmes complexes pouvant comporter jusqu'à plusieurs centaines de contradictions. Dans le but de développer un maximum l'aide informatique apportée pendant la résolution, nous avons construit une ontologie des concepts de la théorie étendue[4] (Cavallucci *et al.* 2009) qui sera appelée Conception Inventive ou CI dans la suite. Cette ontologie va permettre ici, d'une part de représenter un ensemble de brevets traitant du même artefact avec ses évolutions successives comme un ensemble structuré de brevets reliés par des liens avec une sémantique précise, d'autre part d'accéder dans chaque brevet aux connaissances exprimant des concepts pertinents en CI.

Un expert TRIZ, lorsqu'il analyse un problème va construire un modèle de ce problème. La structure de ce modèle dépend de la TRIZ, et contiendra des données qui dépendent du domaine d'application. Ces données seront, en général, fournies par un expert du domaine. Le but de notre outil d'extraction dans les brevets est d'aider à la collecte de ces données dépendant du domaine, et de leur mise en forme dans le modèle du problème. Associé au logiciel TrizAcquisition conçu par le LGECO pour assister l'application de la méthode de CI, cet outil permet donc d'automatiser ou d'aider le travail de deux experts : de CI et du domaine.

Après un bref aperçu de l'état de l'art en section 2, nous donnons en section 3 une explication des principaux concepts de l'ontologie TRIZ nous nous positionnons parmi les principales approches d'extraction de connaissances à partir des brevets. Nous présentons ensuite les tâches à effectuer que nous avons identifiées et les difficultés de chacune d'elles. En

2 Exemple de l'éolienne: P1 = taille des pales ; A = grand; = petit, P2 = puissance générée, P3 = résistance aux vents violents

3 Exemple : une éolienne de taille moyenne serait un compromis; mais une éolienne avec un axe vertical est une évolution inventive, car elle résout la contradiction.

4 La TRIZ n'a jamais été formalisée et l'ontologie que nous proposons est la seule existant à notre connaissance qui précise le sens des termes principaux utilisés par les experts TRIZ et qui permet de formaliser les modèles proposés par cette théorie.

conclusion, nous donnerons des pistes pour réaliser une chaîne de traitement des brevets en vraie grandeur.

1.3. Spécificités de la tâche et difficultés rencontrées

Nous avons choisi de travailler sur des brevets rédigés en anglais plus faciles d'accès. Mais les difficultés rencontrées ne sont pas propres à l'anglais et les méthodes que nous avons choisies sont adaptables à d'autres langues.

Nous avons constitué un corpus initial représentatif de la langue des brevets, qui comprend 100 brevets électroniques publiés entre 2000 et 2009, téléchargeables sur le web (sources *www.googlepatents.com* et *www.patents.com*) appartenant à divers domaines. Nous avons collecté ces brevets en faisant des requêtes avec des mots clés génériques non liés à un domaine particulier. La collection de textes a été convertie en XML, avec une annotation structurelle.

Les principales difficultés de l'accès aux informations contenues dans les brevets viennent de la spécificité de ces textes. Il s'agit de documents relativement courts généralement structurés en paragraphes avec des titres de section standard. Ils sont caractérisés par la présence de phrases longues comportant de nombreuses répétitions et des énumérations dans un jargon très particulier, ni proche de la langue générale, ni de la langue scientifique. Une autre source de difficulté a pour origine le fait que le but visé par le dépôt d'un brevet est plus d'ordre juridique (protection des droits intellectuels) qu'explicatif (explications souvent volontairement confuses). Pour toutes ces raisons, la plupart des outils linguistiques développés pour des textes généraux ou descriptifs, appliqués aux textes de brevets ont de piètres performances.

Par rapport à l'objectif visé ici, un autre problème survient du fait que le modèle abstrait fournit par la CI ne correspond généralement pas à la logique discursive habituellement suivie dans les brevets.

2. État de l'art

S'il existe de nombreuses applications en extraction d'informations et de connaissances, relativement peu sont spécialisées dans les brevets. Un important besoin existe cependant dans ce domaine, car un ingénieur expert doit lire des dizaines, voire des centaines de brevets pour se faire une idée de l'état de l'art dans son domaine d'activité spécifique.

Généralement, les applications spécialisées dans le traitement des brevets utilisent des méthodes hybrides (statistiques–linguistiques). La plupart opèrent un prétraitement linguistique qui comprend en général tokenisation, étiquetage, délimitation des phrases, et généralement reconnaissance des entités nommées. Elles contiennent généralement un module de règles écrites à la main par des experts du domaine soit par des linguistes soit conjointement. Elles utilisent également des méthodes statistiques pour assurer une certaine robustesse. Soit les connaissances humaines peuvent sont exploitées de manière automatique grâce à un module de règles à l'intérieur du système, soit elles servent de base d'apprentissage pour des algorithmes statistiques. Certains systèmes ne visent pas à automatiser complètement le processus : ils font appel à l'intervention extérieure d'un expert qui évalue et trie les résultats fournis par le logiciel.

Une caractéristique importante qui différencie les approches est la nature des informations cherchées: informations de domaine, termes du domaine, informations sur l'évolution de l'artefact. Quant à celles qui s'appuient sur des ontologies, celles-ci ont des rôles divers : ontologies domaines, génériques, structurelles ou non, statiques ou décrivant des connaissances dynamiques. Certains systèmes utilisent une ontologie domaine fournie extérieurement, malheureusement, celle-ci peut alors ne pas être adaptée à la tâche d'extraction. De plus, l'adaptation de la méthode à un nouveau domaine sera parfois problématique, car elle dépendra de l'existence d'une ontologie du nouveau domaine. À partir d'une ontologie, il reste à construire semi automatiquement une Res-source Termino-Ontologique à partir d'un corpus de brevets, afin de relier les concepts aux différentes manières de les exprimer. C'est une tâche concevable, mais qui nécessite du temps et des efforts humains importants . C'est pourquoi nous avons choisi une solution alternative : celle d'automatiser le plus possible la démarche complète, tout en gardant la possibilité de compléter le système par une ressource ontologique.

Nous donnons ici un aperçu sur l'état de l'art dans l'analyse des brevets ainsi que des méthodes existant en extraction de connaissances. Certaines recherches, plutôt linguistiques, par exemple (Guyot *et al.* 2004) se sont intéressées aux caractéristiques qui définissaient le genre des textes des brevets, mais la plupart relèvent du domaine du Traitement Automatique des Langues. (Feldman *et al.* 1998 présentent une approche d'extraction de connaissances à partir de textes non structurés applicable aux brevets. Après un prétraitement linguistique de base, ils produisent d'une liste de mots candidats à être des mots clés. Un filtrage morpho-syntaxique et un filtrage par pertinence statistique sont appliqués à la liste, qui est ensuite

proposée à l'utilisateur pour l'aider à construire la taxonomie du domaine de l'artefact.

Parmi les traitements spécialisés dans les brevets qui s'appuient sur la structure et des propriétés spécifiques du document, il faut mentionner (Ghoula *et al.* 2007) qui présentent une chaîne de traitements réalisant une annotation sémantique automatique des brevets, grâce à une ontologie structurelle et à une ontologie du domaine, dans leur cas la biologie. (Agatonovic *et al.* 2008) utilisent la plate-forme GATE (Cunningham *et al.* 2002) pour l'annotation. L'objectif est de créer un outil robuste et efficace pour pouvoir traiter de grandes quantités de brevets. L'analyse produit également des annotations structurelles et des annotations internes : éléments, unités de mesure et autres types d'entités nommées.

Ces deux travaux définissent une chaîne de traitement robuste, mais aucune n'accède à des connaissances propres au processus d'invention, car elles ne prennent pas en compte les évolutions. Patexpert (Mille *et al.* 2008) est un système commercial qui utilise un système de règles et qui produit un résumé automatique en langue naturelle de la partie "Revendications" (Claims) qui est seule prise en compte. (Sheremetyeva 2003) propose un système hybride (statistique-symbolique). Son système effectue une analyse linguistique par une grammaire de dépendances, basée sur un lexique très riche construit à partir de 1 000 brevets annotés manuellement qui contient : informations morphologiques, structures argumentales des verbes et des rôles thématiques des compléments, fréquence des structures argumentales, classification sémantique. L'idée d'exploiter les propriétés linguistiques des brevets et de se baser sur leur spécificité est intéressante, mais le travail est centré sur les connaissances descriptives.

La plupart des approches apparentées à l'extraction des connaissances, considérées jusqu'ici, sont des approches qui sont centrées sur le contenu du brevet et qui tendent soit à interpréter le contenu par rapport à une ontologie du domaine, soit à accéder aux termes du brevet et à la description statique qu'il contient. Elles visent à faciliter la consultation des brevets par des experts connaissant le domaine de l'artefact. Les approches que nous allons voir ne sont pas centrées sur les connaissances statiques, mais opèrent une sélection d'entités orientées évolution dans les textes de brevets. (Goujon 1999) décrit un système de veille technologique qui utilise la méthode dite de l'exploration contextuelle pour extraire des expressions correspondant à quelques notions qu'elle considère intuitivement comme pertinentes : par exemple *changement*, *utilisation* ou *amélioration*. Elle ne s'appuie sur aucun modèle de connaissances. (Cascini *et al.* 2007) qui s'inspirent de la TRIZ est le travail le plus proche du nôtre

pour ses objectifs et pour son utilisation d'outils de TAL. Leur approche est toutefois distincte de la nôtre. D'une part, elle n'est pas basée sur une ontologie formalisée définissant les concepts utiles, d'autre part, elle est centrée sur le repérage des liaisons (qu'ils appellent "fonctions") entre les éléments de l'artefact. Le résultat du traitement est une représentation du brevet sous forme de triplets (élément, fonction, élément) sélectionnés dans la liste de tous les triplets (Sujet, Verbe, Objet) du texte. Cette représentation, quoique consistante en elle-même, ne satisfait généralement pas les experts. L'extraction des contradictions est à peine abordée dans l'article, la notion de contradiction n'y est pas définie de façon formelle et ne fait pas intervenir d'autres concepts tels : paramètre et valeur (voir section suivante).

3. L'extraction des connaissances de la CI

3.1. L'ontologie

L'ontologie de la Conception Inventive est générique. Elle permet parmi les éléments et les sous-éléments d'un artefact qui sont en général nombreux et agencés de façon complexe, de sélectionner ceux qui entrent en jeu dans une évolution possible. Le résultat visé par le système est donc la population dans un domaine donné de cette ontologie générique pour construire un modèle du domaine, très différent des ontologies domaines statiques habituelles, comportant uniquement des informations concernant des évolutions de paramètres et des conséquences d'évolution de ceux-ci.

Notre approche essaie de suivre la démarche de l'expert en CI lorsqu'il étudie les brevets dans la phase initiale et qu'il repère dans les textes de ceux-ci les **paramètres** qui vont lui servir à mieux cerner son **problème**. Il doit savoir quel est le problème à l'origine de l'invention et quelle **solution partielle** ce brevet propose. Chaque problème est à l'origine d'une ou plusieurs **contradictions** que résout le brevet. La rhétorique sous-jacente aux textes des brevets sert donc à exprimer des informations telles : "Considérant tel artefact, tel défaut s'est révélé, le présent brevet apporte une amélioration qui supprime ce problème." Les concepts de notre ontologie que nous précisons rapidement ci-dessous vont permettre de prendre en compte cette rhétorique.

Plus précisément, un problème exprime les caractéristiques insatisfaisantes d'un système : il est généralement décrit par des expressions de jugement négatif. Un brevet propose une solution partielle à celui-ci, qui est exprimée par l'expression d'un progrès, d'une amélioration. Les éléments sont des composants du système. Les éléments qui nous intéressent sont

ceux qui possèdent des paramètres dont les **valeurs** changent au cours de l'amélioration apportée par le brevet. Il y a deux types de paramètres : les **paramètres d'action** sur lesquels on peut agir et les **paramètres d'évaluation** dont on peut constater le changement de valeur.

En fonction des éléments de l'information à retrouver, notre démarche comporte les étapes suivantes : 1) trouver le problème concernant l'artefact auquel le brevet propose une solution, 2) trouver la solution (partielle) ou l'amélioration apportée par l'invention.

Pour ce faire, il a fallu se faire une idée de la façon dont ces textes expriment les informations en question, c'est-à-dire trouver des régularités entre la structure informationnelle et la structure (morpho)syntaxique du texte. Nous avons ensuite testé des algorithmes d'extraction d'information différents et les avons adaptés. Nous avons finalement conçu une méthode hybride : statistique (filtrage et de l'extraction statistique) combinée avec un module de règles basées sur une analyse linguistique. Dans la suite, nous présentons les tâches à effectuer.

3.2. Filtrage des paragraphes

La contradiction résolue par le produit breveté se formule comme une amélioration par rapport à un état antérieur. Ainsi, nous sommes amenés à chercher des parties de texte qui contiennent une référence au déroulement temporaire d'un changement, de préférence complétée par une ou plusieurs expressions de jugement de valeur (jugement négatif sur les caractéristiques qui constituaient le problème, positif sur les paramètres après l'amélioration apportée, (voir les exemples plus loin). L'opposition entre les propriétés des autres systèmes et celles du nouveau système peut amener à découvrir des contradictions. L'identification des parties de texte susceptibles de contenir des valeurs opposées permettra de filtrer les textes à traiter et ainsi de réduit l'espace de recherche où s'appliqueront des méthodes d'extraction plus ciblées.

Structure du document

Certaines généralisations peuvent être formulées sur la structure des parties pertinentes. L'étude des textes a montré que de nombreuses comparaisons sont explicitées dans les parties appelées *Background* ou *Summary of the Invention* qui constituent souvent des sections séparées, mais malheureusement pas toujours. On y trouve généralement des phrases ou des paragraphes entiers qui détaillent les inconvénients des autres systèmes. Les expressions utilisées dans cette partie du texte sont reprises plus tard lors de la description de l'invention. Les paragraphes pertinents

ne correspondant pas toujours à des sections titrées nous avons conçu une méthode qui filtre ces paragraphes par leur contenu.

Les paragraphes des parties pertinentes s'organisent typiquement autour des thématiques suivantes :

- -a) description du fonctionnement des autres systèmes,
- -b) difficultés, inconvénients ou risques encourus lors de l'usage des autres systèmes,
- -c) objectifs de l'invention, notamment ceux qui visent à éliminer les problèmes mentionnés auparavant,
- -d) description partielle du système breveté, qui explique comment il arrive à résoudre ces problèmes.

Un paragraphe parle, en général, soit des autres systèmes, soit de l'invention et il est très rare de voir des paragraphes dans lesquels il y ait à la fois des phrases qui parlent de l'état de l'art et des phrases qui parlent de l'invention. Nous classifions les paragraphes en trois catégories : 1) ceux qui parlent de l'état de l'art (a-b), 2) ceux qui décrivent la solution apportée par l'invention (c-d), et 3) les autres, c'est-à-dire ceux qui ne sont pas intéressants pour nos objectifs.

Comme les paragraphes de type a) - d) présentent tous des traits linguistiques spécifiques, ils peuvent être repérés automatiquement sur la base de ces spécificités (p. ex. les paragraphes qui parlent de l'état de l'art utilisent le pluriel beaucoup plus fréquemment que les autres parties du texte : "*Most systems are*… "). Ces marqueurs linguistiques caractéristiques ouvrent la voie à plusieurs méthodes pour classifier les paragraphes automatiquement dans une des trois catégories suivantes : État de l'art/Invention/Autre. L'objectif est de filtrer les paragraphes et ainsi réduire la quantité de texte à analyser, tout en préservant le maximum d'informations pertinentes, mais aussi de définir la référence des propriétés énumérées dans le document, de créer le lien entre un élément et un de ses paramètres et ainsi à savoir entre quels sous-systèmes il faut chercher les oppositions qui décrivent la contradiction.

Marqueurs linguistiques

Les régularités cherchées correspondent à des concepts génériques, il n'est donc pas utile de disposer d'une ontologie du domaine. En effet, nous avons observé qu'un nombre relativement restreint de structures linguistiques est utilisé de manière répétitive pour décrire l'état de l'art et

l'apport de l'invention dans le domaine. Voici quelques exemples pour montrer les régularités :

État de l'Art – inconvénients des systèmes "typiques" :

Injection molding is typically done in molds which operate at high temperatures...
Conventionally, the fluid cover stock material enters the mold cavity...
Most injection molds comprise halves that mate to define an internal cavity...
Typical molds include means to heat the molds at numerous point...

However, the known molds of this type still require substantial changeover time…
This presents disadvantages both in cost and in the downtime required to change over a molding machine from one part to another...
Although a two level stack mold can produce product at roughly twice the rate possible with a non-stacked mold, mold costs are considerably higher because of…

Résumé de l'invention – solutions :

A new retractable pin mold for golf balls has now been discovered which alleviates a number of the problems of conventional golf ball retractable pin molds.
An object of this invention is to provide improved quick-changeover cavity inserts…
The present invention also provides increased reliability in the feedback control loop as it enables the user to eliminate numerous junctions which can introduce errors into the control system.

Vu le petit nombre de textes annotés dont nous disposions, nous avons voulu vérifier la généralité de notre algorithme. Nous avons effectué un test sur 12 textes annotés à la main, en associant à chaque paragraphe une des catégories *État de l'art/Invention/Autre*. Nous avons construit des listes de marqueurs établies autour de deux notions : la temporalité (précédent ou typique vs nouveau) et l'amélioration (problème vs solution) :

État de l'art : *typical, conventional, generally, usually, most, often, known*
Invention : *invention, object, relates, provides*
Problème : *disable, damage, disadvantage, loss, error, risk, undesirable, fail, difficulty*
Amélioration : *capable, advantage, can, allow, able, possible, advantageous*

On pourrait se contenter de chercher ces marqueurs dans les paragraphes et appliquer une recherche booléenne donnant ainsi un poids élevé aux marqueurs collectés. Cependant, si la précision des ressources construites à la main est généralement élevée, c'est au détriment du rappel. Or, ces listes sont probablement loin d'être exhaustives, car le corpus qui a servi à la collecte des marqueurs est petit. Dans ce cas, cette méthode donnerait des résultats nettement moins probants sur de nouveaux textes, différents structurellement ou thématiquement du corpus initial. Pour plus de robustesse, nous avons donc choisi d'opérer une classification probabiliste,

où les marqueurs trouvés lors de l'analyse, et chaque mot du paragraphe joueront un rôle dans la classification, mais avec un poids différent.

Classification par apprentissage supervisé

La classification probabiliste a l'avantage de s'appuyer sur un ensemble de propriétés significativement plus grand que celui d'une liste construite à la main. Cependant, pour fonctionner de manière à la fois robuste et précise, une telle méthode a besoin d'un grand corpus d'apprentissage (composé de textes annotés à la main). Comme nous ne disposions que d'un petit corpus de 12 textes annotés selon les trois catégories, nous avons décidé de combiner les deux approches pour obtenir une valeur optimale de rappel et de précision. Nous nous sommes servis des listes de marqueurs extraits à partir du corpus de 12 textes pour construire semi automatiquement un corpus d'entraînement qui fournisse davantage d'exemples positifs (c'est-à-dire de paragraphes classés de manière correcte dans l'un des deux deux groupes pertinents : *Prior Art* ou *Invention*). Une partie du corpus de 100 brevets, contenant 1361 paragraphes, a été annotée à l'aide des marqueurs, et l'annotation a été corrigée à la main. Les marqueurs ont été rangés dans deux catégories suivant leur fiabilité, un poids deux fois plus important a été associé aux marqueurs les plus sûrs qu'aux marqueurs moins fiables.

Par la suite, nous avons défini expérimentalement les paramètres de l'algorithme qui donnent une précision maximale par rapport aux paragraphes annotés semi-automatiquement. Puis, nous avons établi les seuils qui définissent l'appartenance à chaque catégorie en fonction du poids et de la quantité des marqueurs présents dans le paragraphe. Nous avons ensuite annoté le corpus entier en utilisant les paramètres qui ont donné les résultats avec la meilleure précision sur le corpus de test de 1 361 paragraphes.

Le corpus de 100 brevets, malgré sa taille limitée, est suffisant pour servir de corpus d'apprentissage pour la classification probabiliste "bayésienne naïve", que nous avons retenue. Cette méthode utilise comme modèle de probabilité sous-jacent un *modèle de traits indépendants*. Elle peut être appliquée de manière efficace à des tâches de classification supervisée, où l'estimation des valeurs de traits se fait par une estimation de probabilité maximale. Ici, l'espace des traits est construit à partir du vocabulaire des paragraphes pertinents du corpus d'apprentissage et la probabilité de chaque mot du vocabulaire d'appartenir à chacune des catégories est fournie par ce corpus.

Une évaluation de la classification sur les 12 textes annotés manuellement, non inclus dans le corpus d'apprentissage a montré que, comme prévu, les nouveaux paragraphes des textes sont correctement classifiés. Les résultats sont prometteurs malgré la taille du corpus d'apprentissage et le fait qu'il soit annoté semi-automatiquement. Nous avons ainsi opéré une réduction de 70 à 80 % en taille des documents. La précision de la classification est de 68 %, tandis que le rappel moyen est 87 %. Ces résultats sont améliorables avec relativement peu d'effort, en utilisant davantage de textes, en complétant la liste de marqueurs et en réappliquant les règles d'annotation sur le corpus d'apprentissage.

3.3. Extraction des contradictions

Une fois les paragraphes identifiés, il s'agit de rechercher des contradictions dans ceux-ci. Les expressions régulières et les mesures statistiques de pertinence aident à identifier le contexte syntaxique et ainsi à restreindre la liste des candidats. Le plus important est de détecter des oppositions. Or celles-ci peuvent s'exprimer au niveau grammatical ou lexical. Certaines répétitions sont également intéressantes dans la mesure où elles peuvent indiquer un segment de texte pertinent.

Les concepts clés que nous cherchons sont les éléments (les composants du système technique), les paramètres des éléments et les valeurs correspondantes. Dans un système technique, parmi les nombreux éléments composant l'artefact, seuls nous intéressent, ceux subissant un changement. Ce changement l'évolution de l'artefact. Les paramètres ont des valeurs qui peuvent avoir des influences soit positives soit négatives. On remarque très souvent que les trois items du triplet paramètres, valeurs, et éléments sont présents ensemble. Cependant, la contradiction est rarement exprimée dans sa totalité : la plupart des documents n'expriment qu'un changement de valeur du paramètre à la fois (amélioration ou détérioration). Dans ces conditions, notre système ne pourra que signaler à l'utilisateur des contradictions qu'il devra valider et auxquelles il rajoutera éventuellement à la main la partie non explicitée.

Il nous faut donc maintenant, d'une part identifier les trois entités, éléments paramètres valeurs, lier paramètre et élément, paramètre et valeur, d'autre part identifier les deux sens de variations des paramètres pertinents marqués par des oppositions trouvées généralement dans des paragraphes différents.

Repérage des éléments

Pour identifier les éléments pertinents de l'artefact, nous cherchons les entités spécifiques du domaine et plus encore les entités spécifiques du brevet en question. Les mots qui les désignent seront donc significativement plus fréquents dans ce brevet que dans le corpus entier.

Nous constituons des listes de candidats à être des éléments en nous basant sur une analyse de surface et en éliminant, grâce à certaines heuristiques et à des listes d'exclusions, les mots composés qui ne peuvent en être. Nous nous sommes intéressés à la fréquence relative des mots désignant des éléments dans le texte lemmatisé et avons finalement utilisé la mesure **tf-idf,** mesure de pertinence fréquemment utilisée en fouille de textes et extraction d'information. [5]

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$tf - idf_{i,j} = tf_{i,j} \times idf_i$$

où ni,j est le nombre d'occurrences du terme i dans le document j. où le dénominateur est le nombre d'occurrences de tous les termes dans le document j, où |D| dénote le nombre total de documents du corpus.

Cette mesure permet de sélectionner environ une dizaine d'éléments candidats par texte avec les réglages que nous avons choisis (une valeur minimale significativement plus basse pour les noms que pour les autres catégories grammaticales). Nous avons cherché les mots avec une valeur tf-idf élevée qui sont en position de tête d'un groupe nominal[6], ainsi que les adjectifs attributs. Les éléments sont repérés avec un bon rappel, mais la densité de leurs occurrences dans le texte demande davantage de filtrage. Aussi, envisageons-nous dans le futur de compiler le corpus d'une façon différente, en sélectionnant d'abord un ensemble de textes appartenant au domaine de l'artefact. La valeur tf-idf rendra alors davantage d'éléments spécifiques du document et moins d'éléments commun au domaine. Mais,

5 Le tf-idf (de l'anglais term frequency – inverse document frequency) est une méthode de pondération qui permet de quantifier l'importance informationnelle d'un mot dans un ensemble de document, un mot présent partout n'apportant aucune information, un mot présent seulement dans un sous-ensemble de documents permet de caractériser ce sous-ensemble.

6 L'analyse des syntagmes nominaux est effectuée grâce au CRFchunker disponible à http://crfchunker.sourceforge.net/

il restera à voir si cette meilleure sélection n'introduira pas de silence dans le remplissage de notre modèle CI.

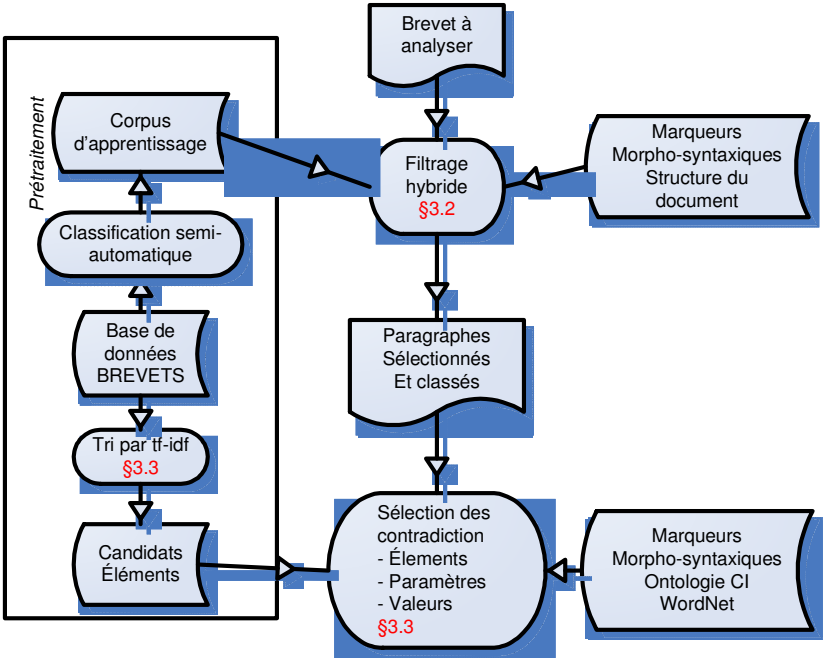

Figure 1. Vue générale de l'agencement des tâches

Module linguistique

Les éléments identifiés par le filtrage statistique doivent de toute façon être filtrés dans une phase ultérieure sur des critères linguistiques. Le module linguistique intégré au système est implémenté comme un ensemble de transducteurs, il permet de créer les liens entre éléments, paramètres et valeurs et de fournir des annotations correspondant aux rôles qu'ils remplissent selon la CI.

Nous avons utilisé plusieurs outils : l'outil LIKES (Rousselot *et al.* 2007) pour retrouver automatiquement les segments répétés, le concordancier Oxford Wordsmith Tools[7] pour sa fonction de tri des concordances, Tree Tagger (Schmid 1994) pour l'étiquetage du corpus.

Après l'analyse linguistique du corpus, nous avons collecté les marqueurs susceptibles d'être des candidats, puis nous avons sélectionné les plus efficaces. NOOJ[8] a servi à les mettre en œuvre pour l'annotation et

7 En vente à Oxford University Press
8 Téléchargeable à http://www.nooj4nlp.net/pages/nooj.html

l'extraction. Nous avons retenu 60 verbes, 137 adverbes, 473 adjectifs et 273 noms. Nous avons constaté que les verbes utilisés dans les brevets sont, dans la plupart des cas, des verbes d'action, plus précisément, des verbes de changement ou des verbes indiquant un changement d'état. Ce sont les plus productifs quant à la détection des valeurs et des paramètres. Les verbes modaux fréquents dans ces textes expriment généralement différents degrés de possibilité (nécessité ou certitude). Lorsque ces modaux sont accompagnés de l'auxiliaire "be" dans le corpus et qu'ils sont suivis de certains indices grammaticaux, ils permettent de localiser certaines informations recherchées.

Les adjectifs sont souvent porteurs de valeurs. Nous avons remarqué également l'usage fréquent des oppositions entre adjectifs dans le corpus.

Les adverbes sont une catégorie généralement difficile à étudier, car le sens de la phrase dépend de leur portée. Cependant, la sélection des adverbes qui nous intéressent a été plus facile, car, ici, seuls nous intéressent les adverbes d'évaluation.

Les rédacteurs de brevets utilisent des noms composés très complexes afin de véhiculer un maximum d'information dans une phrase et ils utilisent un grand nombre de termes pour décrire l'artefact et ses composants. De ce fait, nous ne retenons dans la liste des noms que ceux qui correspondent à des paramètres ou à des valeurs.

Nous avons implémenté les grammaires d'annotation dans NOOJ. Les grammaires correspondent à des transducteurs enrichis (usage de variables et de contraintes, consultation de dictionnaires lors de l'analyse). Nous avons constitué deux dictionnaires spécifiques et édité 46 graphes sur la base des résultats de l'analyse. Les graphes définissent les contraintes à respecter pour effectuer l'annotation. Par exemple, les oppositions doivent être au même endroit ; les verbes doivent être accompagnés de certains indices pour pouvoir être annotés ; l'annotation s'effectue seulement dans le cas où deux notions recherchées au moins existent, etc.

L'application des grammaires d'annotation fournit finalement en sortie le texte annoté en format XML exportable.

Consultation de Wordnet – recherche d'antonymes

Après avoir identifié les éléments qui participent aux changements, il reste encore à chercher les paramètres et les valeurs qui y sont attachées. Un module qui fait appel à Wordnet va essayer de repérer des valeurs opposées, sachant que les oppositions se trouvent entre les descriptions du Prior Art et de l'invention. Les paramètres qui changent de valeur peuvent

s'exprimer d'une part, au niveau lexical, soit par des adjectifs ou participes antonymiques, soit par des paires de verbes : affirmatif – négatif, d'autre part, au niveau syntaxique, par des marqueurs syntaxiques complexes, qui indiquent les rôles joués par les entités situées dans leur contexte proche.

Les oppositions lexicales sont plus faciles à identifier : il s'agit de paires d'adjectifs ou de participes antonymiques, qui sont liés (référentiellement ou syntaxiquement) aux mêmes éléments (syntagmes nominaux). Par exemple :

However, the plastic materials which can be released by resiliently deforming such an undercut area in the prior art injection blow molding process are limited to relatively ***soft*** *plastic materials.*

Another object of the present invention is to provide an injection mold which can release the core mold by resiliently deforming the undercut formed on the lip portion even if it is molded of a relatively ***hard*** *plastic material.*

À part les antonymes, présents à l'intérieur de groupes nominaux ayant une structure identique, nous avons aussi remarqué la présence fréquente de marqueurs d'opposition : par exemple '*limited to*' vs '*even if*' qui permettent d'exprimer des valeurs opposées.

Les antonymies exprimées par des adjectifs à l'intérieur des groupes nominaux ('*hard plastic materials*') ainsi qu'entre les adjectifs ou participes qui ont une fonction prédicative en tant que têtes syntaxiques sont utiles. Il existe également des cas où des substantifs réfèrent à des propriétés exprimant des valeurs de paramètres. Le module fait appel à Wordnet et sélectionne des couples d'adjectifs parmi ceux trouvés dans WordNet, en excluant les adjectifs qui réfèrent à la position, à l'ordre, etc. (p.ex. first-second/last, inner-outer) et qui précisent généralement les éléments du système technique sans exprimer de jugement de valeur.

Les oppositions syntaxiques sont, elles, plus difficiles à localiser. Elles se manifestent souvent par des répétitions lexicales dans des contextes différents, par exemple la même action exprimée une fois dans un contexte affirmatif, et plus tard reprise dans un contexte négatif :

However, such a mold structure disables the release of a molding from the mold. Namely, the undercut of the molded preform as well as the mold will be damaged when the injection core mold is drawn out from the interior of the molded preform.

It is therefore an object of the present invention to provide an injection mold which can injection mold a preform to be biaxially stretch blow molded with a lip portion having an undercut and also which can release the core mold without damaging of the undercut .

Les deux paires d'oppositions syntaxiques sont les suivantes :

disables the release of a molding	*vs*	*can release the core mold*
the undercut will be damaged	*vs*	*without damaging the undercut*

Ces structures peuvent être trouvées par des expressions régulières, comme vu plus haut. Cependant, alors que les éléments et leurs paramètres, ainsi que les paramètres et leurs valeurs sont toujours à chercher dans la même phrase, les oppositions doivent être cherchées dans des paragraphes différents. La recherche doit donc tenir compte des répétitions lexicales et examiner les contextes grammaticaux des segments répétés pour en extraire les oppositions potentiellement pertinentes.

4. Conclusion et directions futures

Nous avons présenté les différentes tâches à effectuer pour extraire des connaissances orientées changement à partir des textes de brevets. Nous avons mis en place une méthode hybride s'appuyant sur des ressources et des outils de traitement de langues pour extraire les informations pertinentes. Notre objectif final est la conception d'un prototype logiciel qui permettra aux inventeurs de connaître l'évolution d'un artefact à un instant donné et comme point de départ pour une future invention.

Le système, encore en cours de développement, comprend des modules de prétraitement linguistique (étiquetage morphologique, analyse syntaxique de surface), un module de fouille de texte statistique, une série de grammaires régulières et, finalement, un module de consultation de WordNet. L'ajout d'un deuxième module linguistique est envisagé pour améliorer les résultats sur le repérage des oppositions. Le contrôle du lancement des différents modules se fait manuellement pour l'instant.

Les études et les expérimentations effectuées ont permis de voir quels éléments incorporer dans la future chaîne de traitement et quelles tâches étaient automatisables ou non.

Elles ont montré qu'il est nécessaire de disposer d'un outil capable de prendre en compte le résultat d'une ou plusieurs expressions régulières et de raisonner sur les contextes dans lesquels on les a trouvés. Pour passer à une plus grande échelle, un module qui facilite l'accès aux bases de données de brevets accessibles sur Internet est également souhaitable.

La chaîne de traitement est en cours de développement autour de LIKES (Rousselot *et al.* 2007) déjà cité. En effet, la toute dernière version de

LIKES possède un plug-in qui permet déjà d'accéder à Google, et bientôt à GooglePatent. LIKES intègre maintenant un système expert basé sur SNARK (Laurière 1986) qui permet d'une part de de faire de la déduction sur le résultat de la recherche d'expressions régulières et d'autre part de lancer des tâches ou des programmes, le système expert servant alors de langage de script. Le système final permettra d'intégrer alors des tâches opérées par des modules Perl (tf-idf par exemple) ou C++(TreeTagger).

Les expérimentations effectuées ont également permis de créer des ressources linguistiques génériques réutilisables, qui devront encore bien sûr être complétées afin d'améliorer la qualité des résultats. Nous savons maintenant que les résultats que nous obtenons sont intéressants, même s'ils doivent parfois être vérifiés et complétés par l'homme. C'est pourquoi, nous projetons d'adjoindre à la chaîne de traitement un module destiné à traiter les liens de références entre brevets et à visualiser ces liens grâce à une interface permettant de cheminer entre eux de manière agréable.

5. Bibliographie

Agatonovic M., Aswani N., Bontcheva K., Cunningham C., Heitz T., Li Y., Roberts I., Tablan V. (2008) : *Large-scale, Parallel Automatic Patent Annotation.* Proc. of the 1st Int. CIKM Workshop on Patent Information Retrieval - PaIR'08, Napa Valley, California, USA

Altshuller Guenrich (1999) : *The Innovation Algorithm* : *TRIZ. Systematic innovation and technical creativity*, Worchester, Mass, Technical Innovation Center

Cascini G. and D. Russo D. (2007) : *Computer-aided analysis of patents and search for TRIZ contradictions.* Int. J. of Product Dev. Vol.4, no.1/2, pp.52-67

Cavallucci Denis, Rousselot François, Zanni-Merk Cecilia (2008) : *Representing and selecting problems in Contradiction Network*, in Proc of the 2nd IFIP Session on Computer-Aided Innovation

Cunningham H., Maynard D., Bontcheva K., Tablan V. (2002) : *GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications.* Proceedings. of ACL'02. Philadelphia

Feldman R., Fresko M., Hirsh H., Aumann Y., Liphstat O., Schler Y., Rajman M. (1998) : *Knowledge Management : A Text Mining Approach.* Proc.of the 2nd Int. Conf. on Pract. Aspects of Knowledge Management, Basel

Ghoula Nizar, Khelif Khaled & Dieng-Kuntz Rose (2007) : *Supporting Patent Mining by using Ontology-based Semantic Annotations*, Proc. of IEEE/WIC/ACM International Conf. on Web Intelligence, Silicon Valley, USA

Goujon Bénédicte (1999) : *Extraction d'informations pour la veille technologique avec le système VIGITEXT*, Actes de RECITAL, Cargese, France

Guyot Brigitte, Normand Sylvie (2004) : *Le document brevet, un passage entre plusieurs mondes*, Actes du Forum pluridisciplinaire document et organisation, Semaine document numérique, La Rochelle

Laurière Jean-Louis (1986) : *Un Langage déclaratif : Snark*, Technique et science informatique, vol. 5, no 3

Mille Simon, Wanner Leo (2008) : *Making Text Resources Accesible to the Reader*, *The Case of Patent Claims*, Proceedings of LREC, Marakesh (Morocco)

Rousselot François, Montessuit Nicolas (2007) : *LIKES un environnement d'ingénierie linguistique et d'ingénierie des connaissances*", Formaliser Les Langues Avec L'ordinateur : De Intex À Nooj", Koeva Svetla, Maurel Denis, Silberztein Max Presses Université de Franche-Comté, Cahiers De La MSH Ledoux ,ISBN 2848671890

Schmid Helmut (1994) : *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Proc. of the Int. Conference on New Methods in Language Processing, pp. 44-49

Sheremetyeva Svetlana (2003) : *Natural Language Analysis of Patent Claims*, Proc. of the ACL-2003 workshop on Patent corpus processing, Sapporo, Japan

Verhaegen P-A., D'hondt J., Vertommen J., Dewulf S., Duflou J.R. (2008) : *Searching for Similar Products through Patent Analysis*. Proc. of the ETRIA TRIZ Future 2008 Conf, Twente

Zanni Cecilia, Cavallucci Denis, Rousselot François (2009) : *An ontological basis for computer aided innovation*, Computers in Industry, ISSN 01663615

A propos des auteurs

Kata Gábor

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
katagood@gmail.com

François Rousselot

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
francois.rousselot@insa-strasbourg.fr

François de Bertrand de Beuvron

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
francoisdebeuvron@insa-strasbourg.fr

Denis Cavallucci

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
dénis.cavallucci@insa-strasbourg.fr

Dildar Wu

LGECO – INSA de Strasbourg
24, bd. De la Victoire
67000 Strasbourg
http://lgeco.insa-strasbourg.fr
angellawooh@gmail.com