CLASSSYN - Classification des documents textuels : La prise en compte du niveau syntaxique

Le projet de recherche CLASSYN se propose d'explorer les potentialités offertes par la prise en compte du niveau syntaxique pour améliorer un système de classification multilingue de documents (français, allemand et anglais), afin d'améliorer les résultats d'un moteur de recherche. Un premier objectif du projet est de développer un système de classification de documents qui combine des techniques classiques statistiques de classification et plusieurs catégories d'information linguistique: lexicales, morphologiques, syntaxiques. Un deuxième objectif concerne la mise en place des ressources nécessaires pour l'outil de classification: un faisceau d'indices morphologiques et syntaxiques pour chaque langue, afin d'obtenir des résultats comparables pour les trois langues étudiés. Un troisième objectif de ce projet est de constituer un corpus multilingue parallèle, annoté syntaxiquement, avec des informations comparables pour les trois langues concernées.

Problématique

Ces dernières décennies, la popularisation de l'ordinateur et le développement des réseaux de communication ont provoqué une explosion du nombre de documents numériques. Ces documents occupent un rôle croissant dans les organisations où ils facilitent les échanges d'information et l'interaction entre les utilisateurs. Vecteurs de culture et de technologie, ils ont une importance stratégique pour nos sociétés. Ils mettent en valeur les archives en permettant aux utilisateurs de se faire assister par l'ordinateur pour la recherche d'information et l'extraction de connaissances. Ces documents numériques couvrent une large palette de formats que les moteurs de recherche avancés doivent être capables de traiter : messages électroniques, articles journalistiques ou scientifiques, flux RSS, blogs, rapports de synthèse, textes de loi, etc.

Devant la quantité vertigineuse d'information, il est devenu indispensable d'améliorer les techniques de recherche d'information et d'extraction de connaissances. De même, il est nécessaire de savoir traiter de manière adaptée les divers formats existants. Les outils disponibles ne fournissent pas encore de résultats complètement satisfaisants. Ils génèrent encore trop souvent du bruit (proposition de documents non souhaités) et du silence (documents pertinents non retrouvés). Les raisons de cette insuffisance résident dans le fait que ces outils négligent généralement les aspects linguistiques du problème. Ils se contentent généralement de rechercher des mots-clés sans prendre en compte des phénomènes comme l'ambiguïté (un mot a plusieurs sens) ou le polymorphisme (un concept peut être exprimé par différents mots). De même, ils s'interrogent rarement sur le contexte d'emploi qui pourrait permettre de lever certaines ambiguïtés ou plus simplement de trier ou de filtrer les résultats. Les moteurs de recherche ne prennent pas en compte des éléments comme la structure du texte, le genre ou le domaine d'application souvent très variés. Les méthodes d'indexation par mots-clés sont appliquées à l'ensemble des documents sans tenir compte de la structure et du type de chaque document. Pourtant, chaque genre textuel ou registre est caractérisé par une série de paramètres linguistiques spécifiques (Biber 1988) : préférence pour un faisceau d'indices comme le temps ou le mode, pour des structures syntaxiques élaborées, etc. Les textes spécifiques à un domaine sont également caractérisés par un lexique spécialisé (Jalam et Cauchat 2002). Un texte argumentatif scientifique n'a pas la même structure qu'un dialogue comportant la présentation d'arguments sur un sujet donné. Un texte spécialisé est caractérisé par un lexique spécialisé et par une organisation textuelle particulière. Même chaque auteur a son propre style, caractérisé par la préférence pour certains constructions

syntaxiques, temps, modes, connecteurs discursifs. L'exploitation de tels paramètres linguistiques est utilisé par les systèmes de classification de textes qui détermine la classe dont appartient un document (selon le domaine, le genre, l'auteur etc.). Chaque document est représenté sous forme de vecteurs de mots, de catégorie lexicale ou d'autres indices linguistiques et une mesure de similarité est appliquée pour déterminer la classe auquel apartient le document. Cette méthode de classification pourrait servir à améliorer les résultats des systèmes de recherche d'information (une amélioration est possible à la fois dans la phase d'indexation et dans la phase de recherche) mais cette approche n'a été que peu exploitée (Manning et al, 2008). Les méthodes de classification qui utilisent des indices linguistiques sont de plus en plus utilisées pour filtrer les résultats et restreindre ainsi le nombre de réponses non pertinentes.

Domaines d'application

Outre l'amélioration des résultats de moteurs de recherche traditionnels, la thématique de notre projet de recherche trouve des applications industrielles dans tous les domaines nécessitant la classification de nombreuses données textuelles ou la sélection de données pertinentes à partir d'une masse importante d'information. Des applications pratiques sont envisageables dans le domaine de la veille technologique ainsi que dans des systèmes d'agrégation de contenu (Mashup) qui permettent d'enrichir des sites Web de manière automatique et pertinente. Ces systèmes d'agrégation de contenu retrouvent les documents liés à une thématique donnée et permettent de mettre à jour régulièrement le contenu des sites Web. Actuellement, il s'agit d'une simple recherche par mots-clés et les résultats sont filtrés manuellement. En effet, dans ces deux cas, la capacité à sélectionner et à classer les informations intéressantes représente un (voire « le ») savoir-faire clef du domaine. Face à la masse importante d'information disponible sur Internet, l'automatisation de ce savoir faire devient une nécessité et les techniques basées sur une simple analyse lexicale donnent des résultats qui ne sont que moyennement satisfaisants. La prise en compte du niveau syntaxique constitue une étape indispensable avant d'aborder le niveau sémantique afin d'améliorer la précision.

La classification pourrait aussi trouver des applications en terminologie et en traductologie où elle peut permettre de trier ou de filtrer les résultats obtenus avec un concordancier afin de mieux cerner le contexte dans lequel un terme est employé. Une autre application serait la différenciation entre textes sources et textes cibles et la reconnaissance éventuelle de paramètres syntaxiques permettant d'identifier la langue source d'un texte traduit.

En linguistique de corpus, la classification permet de partitionner plus facilement des corpus de grande taille.

Le cas particulier qui consiste à identifier un auteur est utilisé en criminologie et notamment dans la lutte contre les terroristes communiquant par l'intermédiaire d'Internet (Zheng et al. 2003, Abbasi & Chen 2006). Dans les domaines de l'éducation et de la publication, ces méthodes sont également utilisées pour détecter le plagiat (Uzuner et al. 2005).

Positionnement

L'objectif de ce projet est d'explorer une méthode de classification de textes qui combine plusieurs catégories d'indices linguistiques pour améliorer les systèmes de recherche d'information.

Face à la globalisation des systèmes d'information produisant un flux multilingue, une première opération de classification est nécessaire pour identifier la langue des documents. L'identification de la langue constitue la première opération à caractère linguistique, c'est elle qui permet de passer du document textuel au document linguistiquement identifié, nécessaire pour pouvoir effectuer toute autre opération.

Une fois la langue connue, des ressources et des outils linguistiques peuvent être utilisés pour améliorer la recherche : par exemple, les mots grammaticaux pourront être éliminés afin de réduire la taille de la table d'indexation ou un outil de lexémisation pourra être utilisé afin d'élargir la recherche aux mots apparentés.

Le processus d'identification de la langue est en fait ni plus ni moins qu'un processus de classification. La classification est effectuée uniquement au niveau morphologique de la langue. Cependant, une fois la langue identifiée une classification de plus haut niveau est possible. Les différentes méthodes de modélisation et les techniques de confrontation sont en grande partie réutilisables pour procéder à une classification selon d'autres critères comme la forme de discours, le genre ou le style.

Ces méthodes de classification qui permettent de trier, de filtrer ou d'identifier sont appréciées depuis longtemps. Cependant, de récentes publications (Abbasi & Chen 2006) montrent que l'aspect linguistique est encore peu exploité. Alors que les auteurs disent que leur méthode de classification est basée sur des caractéristiques syntaxiques du texte, il s'avère qu'ils ne prennent en compte que la fréquence d'occurrence des différentes ponctuations et de quelques mots grammaticaux.

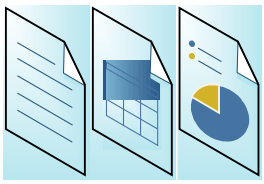
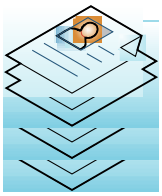
Cette manière de procéder est effectivement la plus simple puisqu'elle ne nécessite aucune analyse syntaxique préalable. Les méthodes de classification actuelles se limitent à regrouper les documents par termes communs ou par sujet (Cavnar et Trenkle 1994, Joachims 1998). La classification identifie dans les documents des caractéristiques qui sont généralement des mots ou des séquences de caractères plus courtes et appliquent une mesure de similarité entre les documents pour détecter des classes de documents similaires. Peu de paramètres linguistiques sont pris en compte par les systèmes de classification automatique. Malgré les travaux de recherche menés sur l'identification automatique du genre textuel (Poudat 2006, Haber et al. 2000), les méthodes restent applicables à un domaine particulier ou à des catégories de texte bien définies (articles scientifiques). Certaines de ces méthodes se concentrent sur une typologie d'indicateurs pour identifier les séquences argumentatives (Teufel et Moens 1999, Buckingham Shum 2007). Peu de moteurs de recherche ou d'outils d'extraction de connaissances font appel à un profilage des textes même si certains travaux proposent une classification des genres sur Internet (Baroni et Bernardini 2004).

Il nous semble néanmoins que la prise en compte de réelles propriétés syntaxiques devrait apporter une plus-value non négligeable. Nous proposons donc de mettre au point un classificateur prenant en compte de réelles propriétés syntaxiques. Pour identifier le genre textuel, nous utiliserons des paramètres linguistiques inspirés des travaux de linguistique textuelle (Adam 1999, Bakhtin 1986, Ducrot et al. 1980) et des méthodes statistiques. Nous envisageons le développement d'une méthode incrémentale permettant de prendre en compte les différents niveaux de paramètres syntaxiques : chunks, constituants syntaxiques, dépendances, etc.

Objectifs et débouchés industriels

Nous proposons de faire notre travail de recherche en partenariat avec l'entreprise GreenIvory, implantée en Alsace (à Haguenau). Notre travail de recherche trouvera en effet une application industrielle à travers l'intégration du classificateur de textes dans MashupXFeed, un des produits stratégiques de l'entreprise GreenIvory. Cette entreprise est spécialiste du Mashup (systèmes d'agrégation de contenu permettant d'enrichir automatiquement les sites Web) et travaille en outre sur plusieurs projets logiciels de veille et d'e-réputation (fouille automatique d'opinions sur un produit ou une organisation). MashupXFeed est un produit qui lance des recherches d'informations sur le Web pour retrouver des informations récentes sur un sujet donné. Les documents retrouvés sont ensuite filtrés manuellement, afin de mettre à jour le contenu d'un site Web. Le classificateur permettra l'automatisation de cette étape de filtrage (2ème étape du tableau ci-dessous).
La coopération avec l'entreprise nous donnera en outre l'opportunité de confronter nos résultats aux besoins du monde industriel.

Les 4 étapes du mashup :

Etape	Description	
1 Agrégation	Lecture des différentes sources d'information (flux RSS, e-mail, etc.).	YAHOO! Google DNA DERNIERES NOUVELLES D'ALSACE …
2 Filtrage	Nettoyage de l'information : élimination des concurrents, des indésirables, détection de la date, formatage des données…	
3 Amélioration	Détection des éléments d'information, recherche d'information complémentaire, création et formatage de l'information grâce à la technologie GreenIvory **MashupOverlay**.	
4 Publication	Publication de l'information sur un portail, sur un site dédié, dans une newsletter ou un e-mail ou encore de façon interactive.	

Méthodologie

Comme un des objectifs du projet est d'évaluer l'apport que l'information syntaxique pourra avoir pour une amélioration d'un système de classification automatique, nous proposons d'adopter la méthodologie suivante :

- Identification des corpus crées par l'équipe ou disponibles, étiquetés et annotés (Acquis Communautaire, Hansard, Europarl) et constitution des corpus multilingues parallèles (manuels techniques dans le domaine de l'automobile et des manuels d'utilisation de logiciels) et étiquetage morphosyntaxique pour les trois langues ;
- comparaison des outils qui proposent une analyse syntaxique détaillée (VISL (Bick 2003) disponible pour le français, l'allemand et l'anglais ; FIPS (Wehrli 2007), XIP (Xerox Incremental Parser) (Aït-Mokhtar, Chanod et Roux 2002), Dyalog (Villemonte de la Clergerie 2002) disponibles pour le français ; la solution de Michael Schiehlen (Schiehlen 2003) pour l'allemand.
- annotation des corpus à l'aide des outils d'analyse syntaxique identifiés.
- identification d'un faisceau optimal de paramètres linguistiques permettant de décrire la forme de discours ou le genre. D'un côté, nous allons étudier le lexique et des unités phraséologiques comme les collocations (Dini et al. 1998), (Todirascu et al. 2008), (Heid, 2008). D'un autre côté, nous pouvons prendre en compte les propriétés morpho-syntaxiques (temps, mode, etc.), les structures syntaxiques ou les relations de dépendance comme critères de classification des documents.
- définition d'une typologie des critères de classification (langue, forme de discours, genre, style, etc.) en concentrant nos recherches sur les études portant sur les langues choisies (français, allemand et éventuellement anglais)
- développement du système de classification de documents hybride.

évaluation de l'apport des indices syntaxiques pour l'amélioration du système de classification.

Nous appuyant sur le travail effectué par (de la Porte des Vaux, 2008) qui portait sur l'identification automatique des langues, nous continuerons à explorer les techniques de classification et nous nous appliquerons plus particulièrement à établir un état de l'art des méthodes utilisées pour la création de modèles autres que des modèles de langue. Nous nous intéresserons particulièrement aux méthodes permettant d'identifier un auteur. Celles-ci font en effet appel à une grande variété d'indices tout en restant indépendantes du sens du texte. Elles s'intéressent à toutes les propriétés de la langue, des caractéristiques morphologiques jusqu'à, dans une moindre mesure, la syntaxe qui fera l'objet d'études plus approfondies de notre part.

Nous réunirons un corpus de textes représentatifs de ces critères et nous permettant d'évaluer les performances du classificateur. Dans la mesure du possible, nous constituerons un corpus parallèle afin de faciliter l'étude contrastive entre les langues. Nous regrouperons des textes de lois (Acquis communautaire), des débats parlementaires (Europarl, Hansard) voire des manuels techniques (constructeurs automobiles, éditeurs de logiciels, etc.). Des corpus comparables pourront être établis sur la base d'articles journalistiques (disponibles dans les trois langues).

Le processus d'identification comprend deux phases principales. La première est une phase de modélisation pendant laquelle le système est entraîné pour pouvoir reconnaître les langues choisies. L'entraînement a pour but de créer un modèle pour chaque langue à partir de textes dont la langue est connue. Cette opération consiste à recueillir des statistiques pour un certain nombre de caractéristiques du texte. Il s'agit d'une phase préparatoire. La seconde phase est la phase de confrontation durant laquelle un texte dont la langue n'est pas connue est confronté à

chaque modèle dans le but de déterminer la langue la plus probable. Le modèle donnant les meilleurs résultats est vraisemblablement celui qui correspond à la langue du texte. Il s'agit d'une phase exécutoire qui permet l'identification proprement dite de la langue.

Dans le cas de l'identification de la langue, les caractéristiques sont des n-grammes de caractères ou des mots grammaticaux. Pour une classification de textes selon d'autres critères que la langue, la méthode classique identifie un ensemble de mots clés dans le document. Le document est représenté sous forme d'un vecteur de caractéristiques et une mesure de similarité est utilisée pour trouver la classe la plus pertinente. Nous proposons d'utiliser d'autres caractéristiques linguistiques comme les étiquettes, les lemmes, les propriétés morpho-syntaxiques (temps, mode, etc.) ou des connecteurs de discours.

Pour cela, nous devrons nous pencher sur les problèmes de l'étiquetage morphosyntaxique et de l'analyse syntaxique. En effet, notre classificateur aura besoin d'informations syntaxiques attachées au texte. Par conséquent, il nous faudra rechercher les outils existants qui proposent une analyse syntaxique détaillée (VISL (Bick 2003) disponible pour le français, l'allemand et l'anglais ; FIPS (Wehrli 2007), XIP (Xerox Incremental Parser) (Aït-Mokhtar, Chanod et Roux 2002), Dyalog (Villemonte de la Clergerie 2002) disponibles pour le français ; la solution de Michael Schiehlen (Schiehlen 2003) pour l'allemand.

Nous évaluerons ces outils et étudierons leur niveau d'analyse ainsi que la manière dont les informations syntaxiques sont codées afin de voir celui se prêtant le mieux à une intégration dans notre classificateur. Nous annotons les corpus avec les outils choisies à partir de cette évaluation.

Nous proposons une typologie des critères de classification (langue, forme de discours, genre, style, etc.) en concentrant nos recherches sur les études portant sur les langues choisies (français, allemand et anglais). Il nous faudra trouver le faisceau optimal de paramètres linguistiques permettant de décrire la forme de discours ou le genre. D'un côté, nous allons étudier le lexique et des unités phraséologiques comme les collocations (Dini et al. 1998), (Todirascu et al. 2008), (Heid, 2008). D'un autre côté, nous pouvons prendre en compte les propriétés morpho-syntaxiques (temps, mode, etc.), les structures syntaxiques ou les relations de dépendance comme critères de classification des documents.

Après avoir choisi les critères de classification nous offrant les perspectives de recherche les plus intéressantes, nous élaborerons ensuite un classificateur en veillant à ce qu'il puisse être intégré facilement dans une chaîne de traitements. Il pourra ainsi être intégré dans un moteur de recherche ou un concordancier pour filtrer ou trier les résultats.

Le logiciel résultant fera l'objet d'une évaluation approfondie qui devra mettre en lumière le degré d'amélioration dû à la prise en compte du niveau syntaxique. En outre, nous étudierons les variations de performance en fonction de la profondeur de l'analyse syntaxique. Les résultats obtenus pour les différentes langues seront analysés dans une optique contrastive.

Plan de l'étude

Pour mener à bien le projet, nous proposons de respecter les étapes suivantes :

- constitution des corpus et étiquetage morphosyntaxique: janvier 2010-avril 2010
- comparaison des outils d'analyse syntaxique pour les trois langues: mai 2010 – août 2010
- annotation des corpus annotés: septembre 2010 – décembre 2010
- identification des paramètres linguistiques: janvier 2011 – mars 2011

- développement du système hybride de classification de documents : mars 2011 – septembre 2011
- évaluation de l'apport des indices syntaxiques pour l'amélioration du système de classification octobre 2011 – décembre 2011

Une partie de l'évaluation de l'analyseur et la phase de constitution du corpus seront effectués à l'IMS de Stuttgart disposant de ressources importantes. De même, l'élaboration du classificateur sera effectuée à l'institut afin de bénéficier au maximum des compétences statistiques et informatiques locales. La majeure partie des autres étapes sera effectuée à l'Université de Strasbourg où nous tenterons de nous inspirer des travaux de Charles Muller[1] (Muller 1979, 1985, 1992).

[1] Ancien professeur de l'université de Strasbourg, passionné de statistiques, il aurait eu 100 ans cette année.

Compétences des partenaires

Pour mener à bien ce programme nous bénéficierons des compétences complémentaires de partenaires provenant aussi bien du milieu académique (France et Allemagne) que du milieu industriel (France) : l équipe LiLPa (Linguistique, Langues et Parole, EA1339) de l'Université de Strasbourg, l'Institut pour le traitement automatique du langage (IMS, Institut für maschinelle Sprachverarbeitung) de l'Université de Stuttgart (Allemagne) et la société GreenIvory (présente en France et aux Etats-Unis). La contribution des partenaires académiques (LiLPa et IMS) sera essentiellement de nature théorique, méthodologique et technique tandis que celle de GreenIvory portera sur les aspects relatifs à l'intégration et la mise en œuvre des travaux dans un contexte industriel.

L'équipe LiLPa (Linguistique, Langues et Parole) est une équipe d'accueil de l'Université de Strasbourg qui regroupe des chercheurs en linguistique, sociolinguistique, didactique des langues, Traitement Automatique des Langues (TAL), linguistique de corpus et phonétique répartis en plusieurs composantes : SCOLIA (Sciences cognitives, linguistique, intelligence artificielle), l'Institut de Phonétique de Strasbourg (IPS), Fonctionnements discursifs, Didactique des langues et le Groupement d'études sur le plurilinguisme européen (GEPE). L'équipe compte 66 enseignants chercheurs et 86 doctorants. Les thèmes de recherche de l'équipe sont : la problématique du sens sous toutes ses facettes (composante SCOLIA), les recherches disciplinaires en didactique de la langue et de la littérature française (Didactique des langues), les recherches en linguistique de corpus, TAL et didactique des langues vivantes étrangères (composante Fonctionnements discursifs), la phonétique (IPS), les politiques linguistiques, le contact de langues et des cultures ainsi que les outils de traduction (GEPE).

La composante la plus impliquée dans le présent projet de recherche est celle travaillant sur les Fonctionnements Discursifs. Cette composante rassemble principalement des enseignants-chercheurs en linguistique des langues vivantes étrangères, en TAL et en lexicologie. Ses activités de recherche spécifiques portent plus particulièrement sur l'étude des variétés en langue et en discours et sur les phénomènes d'implicitation observables à différents niveaux d'analyse dans les langues naturelles, à partir :

- de l'étude des mécanismes de la compréhension, notamment orale,
- de l'analyse des rapports dans le couple image/parole
- des phénomènes de cohésion discursive.

Cette composante comprend un axe de recherche « Lexicologie et TAL » qui travaille sur la phraséologie (expressions figées, collocations,…) et la terminologie en s'appuyant sur une analyse de corpus dans un contexte multilingue. Cet axe étudie aussi la structure thématique des documents afin d'extraire des informations pertinentes de ces derniers. Les résultats de ces études sont utilisés pour la création des ressources linguistiques en format électronique (lexiques, grammaires, corpus annotés), la création des outils TAL pour l'acquisition des connaissances et pour la recherche d'information. Ces activités de recherche se déroulent dans le cadre des projets financés ("Collocations en contexte", projet financé par l'Agence Universitaire pour la francophonie; collaboration industrielle concernant le développement d'un moteur de recherche thématique). Le présent projet s'inspirera des travaux récents de l'équipe (Todirascu et al., 2008 ; Lortal et al., 2007 ; Todirascu, 2006) et utilisera des outils et des corpus développés au sein de cette composante.

Une autre composante LILPA, la composante GEPE (Groupement d'études sur le plurilinguisme européen), est impliquée dans le projet par son axe « Outils de traduction ». Cet axe regroupe des enseignants-chercheurs travaillant sur des questions liées au domaine de la traduction. Ses activités de recherche sont conduites selon trois approches

complémentaires, d'un point de vue aussi bien théorique que soucieux des applications pratiques envisageables :

- didactique : formalisation de la pratique traductionnelle par l'interaction pratique théorie et théorie pratique, formalisation d'une didactique de la traduction par rapport à certains domaines précis comme par exemple le sous-titrage de films.
- traductologie : la qualité en traduction (évaluation et révision) constitue un aspect très important, il advient de dégager des outils théoriques et pratiques qui permettent de la mesurer.
- traitement automatique du langage naturel : la constitution des ressources nécessaires pour la traduction (les dictionnaires électroniques multilingues, l'alignement de corpus et les ontologies).

Travaillant entre autres sur le français, l'allemand et l'anglais, l'axe a adopté une méthodologie d'analyse linguistique contrastive des phénomènes qui posent des problèmes en traduction et qui doivent être pris en compte pour constituer des ressources électroniques efficaces. L'apport de cette composante est principalement un apport méthodologique et technique (Grass et al., 2002 ; Grass et al., 2004).

Le département "Fondements de la linguistique computationnelle" de l'IMS (Institut für maschinelle Sprachverarbeitung, Université de Stuttgart) est composé de 5 collaborateurs permanents, ainsi que d'une dizaine de chercheurs à temps partiel. Les projets de recherche menés par le département s'inscrivent dans le domaine de la linguistique de corpus et de la traduction automatique statistique et vise l'extraction de ressources électroniques pour l'analyse syntaxique automatique à partir des corpus. Le département a développé plusieurs ressources électroniques (corpus étiquetés et annotés syntaxiquement, analyseurs syntaxiques pour l'allemand). L'apport du laboratoire est théorique et méthodologique (Heid, 2008 ; Lapshinova et Heid, 2007 ; Manning, Raghavan et Schütze, 2008) ainsi que technique, par la mise à disposition des outils d'analyse automatique pour l'allemand (étiqueteur, analyseur syntaxique).

Le partenaire industriel de notre projet, la société GreenIvory, fondée en janvier 2006 aux Etats-Unis par un entrepreneur français, Jean Georges Perrin, GreenIvory est présente en France depuis 2007 sous la forme d'une SARL. Cette petite société à vocation internationale est spécialisée dans le traitement de l'information et sa mise en valeur sur Internet. Son cœur de métier est la dynamisation de sites Web de manière automatique. Maîtrisant les nouvelles technologies Internet, elle s'est spécialisée dans le mashup, c'est-à-dire dans les systèmes d'agrégation de contenu permettant d'enrichir automatiquement les sites Web.
L'entreprise française, basée à Haguenau, compte actuellement une petite dizaine de personnes et prend en charge toute la partie recherche et développement. L'équipe R&D, en croissance régulière, est composée d'ingénieurs expérimentés. Elle est menée par Jean Georges Perrin qui s'occupe aussi de la direction stratégique de la société. Ce dernier est ingénieur en informatique avec plus de 12 ans d'expériences professionnelles, dont plus de 8 en gestion de projets innovants, dans des environnements internationaux et hétérogènes. Il a une profonde culture « nouvelles technologies » qui remonte à 1994.
Les technologies de mashup ont pour GreenIvory une importance stratégique. Sa technologie fondamentale réside dans sa capacité à agréger du contenu, l'interpréter et l'enrichir: c'est un processus de mashup. Le processus passe par une recherche d'information sur le Web sur un sujet donné, un filtrage des documents et des informations pertinentes et une intégration de ces informations pour actualiser le contenu des sites Web. Cette technologie est déclinée dans plusieurs produits. Les produits dérivés sont:

- MashupXFeed et MashupXpress qui utilisent cette technologie pour dynamiser des sites.
- VoiceObserver qui utilise cette technologie pour mesurer l'e-réputation d'un produit, d'une marque, d'une personne.

Les produits ci-dessus alimentent et surveillent les réseaux sociaux avec EnterpriseSocialNetwork.

Les travaux de recherche permettrons d'améliorer la solution logicielle MashupXFeed, commercialisée par GreenIvory. Il s'agit d'une suite de composants qui permettent d'enrichir automatiquement le contenu des sites Web. L'enrichissement est obtenu en collectant des données sur Internet, en les filtrant puis en les re-publiant sur le site du client. Des techniques d'ingénierie linguistique sont utilisables principalement au niveau du filtrage des données, tout l'enjeu étant de n'afficher sur le site du client que des informations pertinentes. C'est à ce niveau de MashUpXfeed que le classificateur de textes sera intégré.

Références

ABBASI Ahmed, CHEN Hsinchun, (2006), **Visualizing Authorship for Identification**, Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics (ISI 2006), San Diego, CA, USA

ADAM Jean-Michel, (1999), **Linguistique textuelle. Des genres de discours aux textes**, Paris, France, Nathan

AÏT-MOKHTAR S., CHANOD J.-P., ROUX C., (2002), **Robustness beyond shallowness: Incremental Deep Parsing**, Special Issue of the Natural Language Engineering Journal on Robust Methods in Analysis of Natural Language Data, Cambridge, pp. 121-144

BAKHTIN Mikhail Mikhailovitch, (1986), **Speech Genres and Other Late Essays**, Austin, University of Texas Press

BARONI M., BERNARDINI S., (2004), **BootCaT: Bootstrapping Corpora and Terms from the web**, Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp. 1313-1316

BIBER Douglas, (1988), **Variation across speech and writing**, Cambridge, England, Cambridge University Press

BICK Eckhard, (2003), **A CG & PSG Hybrid Approach to Automatic Corpus Annotation**, SproLaC, Lancaster, UK

BUCKINGHAM SHUM Simon J., UREN Victoria, GANGMIN Li, SERENO Bertrand, MANCINI Clara, (2007), **Modeling naturalistic argumentationin research literatures: Representation and interaction design issues:**, International Journal of Intelligent Systems, pp. 17-47

CAVNAR William B., TRENKLE John M., (1994), **N-Gram-Based Text Categorization**, Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US, pp. 161-175

De la Porte des Vaux, B. (2008). Un système de détection automatique des langues. Mémoire de Master, Université Marc Bloch de Strasbourg.

DINI L., DI TOMASO V., SECOND F., (1998), **Word Sense Disambiguation with Functional Relations**, Language Resource and Evaluation Conference, Granada

DUCROT Oswald et al., (1980), **Les mots du discours**, Paris, France, Editions de Minuit

GRASS, T., MAUREL, D., PITON, O. (2002), **Description of a multilingual database of proper names**, in Ranchod, E. & Mamede, N. (éds.) Advances in Natural Language Processing, Lecture Notes in Artificial Intelligence 2389, Berlin, Springer Verlag, pp. 137-140.

GRASS, T., TRAN, M., MAUREL, D. (2004), **Prolexbase : Une ontologie pour le traitement multilingue des noms propres**, in Rita Temmerman & Uus Knops (éds.) The translation of domain specific languages and multilingual terminology management, Linguistica Antverpiensia 3/2004, pp. 293-309.

HABER B., ILLOUZ G., LAFON P., FLEURY S., FOLCH H., HEIDEN S., PRÉVOST S., (2000), **Profilage de textes: cadre de travail et expérience**, JADT 2000: 5èmes Journées Internationales d'Analyse Statistique des Données Textuelles, Lausanne

HEID, U., (2008), **Computational Phraseology: an overview**, in Sylviane Granger and Fanny Meunier, editors, Phraseology -- An interdisciplinary perspective Amsterdam/Philadelphia: John Benjamins (à paraitre)

HUNNISETT David, TEAHAN William J., (2004), **Context-based method for text categorisation**, Proceedings of the 27th Annual International ACM SIGR Conference (SIGIR), Sheffield, UK

JALAM Radwan, CAUCHAT Jean-Hugues, (2002), **Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques**, JADT 2002: 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St Malo, France

JOACHIMS Thorsten, (1998), **Text categorization with support vector machines: Learning with many relevant features**, Proceedings of the European Conference on Machine Learning, Springer-Verlag

LAPSHINOVA, E., HEID, U., (2007), **Syntactic subcategorization of noun+verb multiwords: description, classification and extraction from text corpora**, in C. Camugli; M. Constant and A. Dister, editors, Actes du 26e Colloque international Lexique et Grammaire, Université Marne-la-Vallée, Bonifacio (Corse), France, pp. 73-80.

LORTAL Gaëlle, TODIRASCU-COURTIER Amalia, LEWKOWICZ, Myriam, (2007), (15 pages) **AnT&CoW: Share, Classify and Elaborate Documents by means of Annotation**, Journal of Digital Information Management, eds. Richard Chbeir, Ajith Abraham, Pit Pichappan

MANNING, Christopher, RAGHAVAN, Prabhakar, SCHÜTZE, Hinrich (2008), **Introduction to Information Retrieval**, Cambridge University Press (à paraitre)

MANNING, Christopher, Schütze, Hinrich (1999). Foundations of Statistical Natural Language Processing. MIT Press, Boston, MA

MULLER Charles, (1979), **Langue française et linguistique quantitative - Recueil d'articles**, Genève, Suisse, Editions Slatkine

MULLER Charles, (1985), **Langue française, linguistique quantitative, informatique**, Genève, Suisse, Editions Slatkine

MULLER Charles, (1992), **Initiation aux méthodes de la statistique linguistique**, Paris, France, Editions Champion

MULLER Charles, (1992), **Principes et méthodes de statistique lexicale**, Paris, France, Editions Champion

POUDAT Céline, (2006), **Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres**, Orléans, France, Thèse de l'université d'Orléans

SCHIEHLEN Michael, (2003), **Combining Deep and Shallow Approaches in Parsing German**, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp. 112-119

TEUFEL S., MOENS M., (1999), **Discourse-level argumentation in scientific articles: human and automatic annotation**, Toward Standards and Tools for Discourse Tagging, ACL 1999 Workshop

TODIRASCU-COURTIER, Amalia, (2006), **Outils TAL pour les textes narratifs**, in SOULIER, Eddie (ed.): Traité IC2 : Storytelling : Concepts, outils, applications, Ed. Hermès, pp. 314 – 336

TODIRASCU, A., HEID, U., STEFANESCU, D., TUFIS, D., GLEDHILL, C., WELLER M., ROUSSELOT F., (2008), **Vers un dictionnaire de collocations multilingue**, Cahiers de Linguistique, Université de Louvain.

UZUNER Özlem, KATZ Boris, NAHNSEN Thade, (2005), **Using Syntactic Information to Identify Plagiarism**, 2nd Workshop on Building Educational Applications using NLP

VILLEMONTE DE LA CLERGERIE Eric, (2002), **Construire des analyseurs avec DyALog**, TALN 2002, Nancy, France

WEHRLI Eric, (2007), **Fips, a Deep Linguistic Multilingual Parser**, ACL 2007 Workshop on Deep Linguistic Parsing, Prague, Czech Republic, pp. 120-127

ZHENG R., QIN Y., HUANG Z., CHEN H., (2003), **Authorship analysis in cybercrime Investigation**, 1st NSF/NIJ, ISI2003, Springer-Verlag, pp. 59-73