Mots-clé : traitement automatique des langues, données légales, système d'information, logique.

1 - Présentation du projet de recherche

A - Situation du projet

Dans un contexte de volume croissant de données et de connaissance, à la fois en quantité et en diversité, un des objectifs de l'équipe SemLIS (groupe porteur du projet) (« *Semantics, Logics, Information Systems for Data-User Interaction* ») est de **redonner du pouvoir à l'utilisateur**. Par ce terme nous entendons un individu ou un groupe avec un intérêt fort pour certaines données (personnelles ou collectives), et le besoin de les exploiter pour en déduire de nouvelles connaissances ou prendre des décisions.

Dans le domaine du droit, il s'agit d'envisager en particulier des systèmes d'aide à un usager et une chaîne de traitements allant des textes bruts ou partiellement structurés, à des représentations formelles et à facettes logico-sémantiques, pour traiter et valoriser des données, en lien avec des règles du droit.

Ce type de projet requiert des compétences complémentaires, dans les domaines suivants : droit, linguistique computationnelle et informatique ; et plus précisément pour ce dernier: génie logiciel, logique, traitement automatique des langues naturelles et fouille de données.

B – Jeu, de données, cas d'étude

Divers cas d'étude peuvent être envisagés. Comme sous-domaine, nous pourrons considérer celui des élections, en collaboration avec d'autres équipes ; ce sous-domaine a pour avantage d'être régi par un corpus juridique réduit. Le respect de certains articles de loi peut être vérifié par l'analyse conjointe de données hétérogènes (exprimées sous forme numériques ou en langage naturel).

Un autre type d'analyse concerne aussi les règles (la procédure d'une élection par exemple).

C - Système d'information, portant sur des données du droit

La vision de l'équipe SemLIS est plus celle d'une collaboration homme-machine que d'une automatisation complète, pour l'expert comme l'utilisateur sans connaissances a priori, confronté à une masse d'informations.

Un objectif essentiel des LIS (Systèmes d'information logique) est de combiner des logiques et des étapes d'explorations pour briser des limitations de systèmes existants pour retrouver et mettre à jour l'information (hiérarchies, recherches booléennes). Cette approche a aussi été appliquée à plusieurs domaines dont des données linguistiques.

Un aspect distinctif de l'équipe est l'application de méthodes formelles provenant du génie logiciel, d'informatique théorique (grammaires formelles, logique, théorie des types, langages déclaratifs, preuves) à des tâches d'intelligence artificielle (représentation des données et raisonnement, extraction de données, interaction).

Dans ce projet, notre champ d'étude sera celui des données à caractère règlementaires, sous diverses formes : textes bruts, semi-formelles ou formelles et logiques. .

2- Méthodologie

A - Spécificité des textes et type d'analyse

Une première spécificité à prendre en compte est la langue utilisée. De nombreux travaux existent pour l'anglais. À notre connaissance peu de travaux concernent le traitement automatique des langues dans le domaine du droit en français. Des approches et des outils généraux existent, mais il faudrait les adapter à plusieurs niveaux :

- terminologie ;

- analyse syntaxique ;

- analyse sémantique ;

- marqueurs du discours.

À ces niveaux correspondent des ressources, des formalismes et des outils informatiques pour représenter et analyser des connaissances du domaine et l'information contenue dans un texte. Mais de ce point de vue, les textes du domaine du droit ont certaines particularités, allant de la terminologie à la structure des textes. Améliorer la couverture des phénomènes linguistiques spécifiques au français et au domaine du droit, intégrer une ontologie du domaine sont aussi souhaitables. Ces analyses automatisées devraient faciliter l'exploitation informatique de ces textes, fournir des représentations utiles et pertinentes, facilitant leur exploration (recherche d'information) et leur maitrise.

Des travaux en logique (et philosophie) peuvent être utiles dans ce cadre, pour les appliquer notamment à certains types de textes présentant des règles ou des argumentations.

B - Approche SemLIS

Nous décrivons dans cette section une des approches possibles pour la partie système d'information. L'approche LIS intègre naturellement des facettes sémantiques multiples et permet en cela d'appréhender deux problèmes essentiels en qualité des données : celui de l'hétérogénéité des données et celui de la sémantique.

Redonner du pouvoir à l'utilisateur se traduit par plusieurs objectifs :

O1: rendre l'utilisateur autonome et agile en évitant des intermédiaires (e.g., administrateur de base de données) pour exploiter les données et la connaissance ;

O2: faciliter l'interconnexion de données hétérogènes et multi-sources ;

O3: apporter de la flexibilité en autorisant l'acquisition de données hors schéma et l'évolution continue d'un schéma de données ;

O4: apporter un niveau de contrôle et de confiance dans le système en favorisant la transparence et la prédictabilité d'un système d'actions ;

O5: permettre une acquisition collaborative et la vérification des données et des connaissances.

Ces objectifs sous-tendent des défis concernant l'extraction d'information (structurée par exemple avec RDF), l'expressivité, la représentation des connaissances, l'interaction données-utilisateur.

L'équipe SemLIS a développé des outils : des systèmes de gestion de contexte LIS. Les données y sont caractérisées par des propriétés logiques permettant une exploration riche et sans connaissance a priori. L'utilisation du système peut être orienté pour la tâche du repérage d'erreurs et d'incohérences (et la proposition d'amélioration). Un autre avantage de l'approche LIS est celle de permettre une forme de sérendipité.

Un autre facteur de qualité concerne les données inconnues ou incomplètes : l'approche LIS fonctionne justement dans ce cadre-là.

Les algorithmes de calcul employés sont principalement issus de l'analyse de concepts formels, de la logique et du traitement automatique des langues.

Nous proposons d'exploiter les données en utilisant les méthodes de l'analyse de concepts logiques [Ferré et Ridoux (2003)] via des outils de l'équipe LIS : Camelis http://www.irisa.fr/LIS/ferre/camelis/ ou Sparklis http://www.irisa.fr/LIS/ferre/sparklis/.

Des expérimentations avec le système de gestion de contexte et le contexte obtenu permettront d'améliorer l'approche et la construction des facettes sémantiques à mettre en avant.