Génération de récits à partir d'expériences spatiales d'un robot de service par extraction de connaissances textuelles

Table de matières

1. Résumé du projet

Le but du projet est de donner à un robot la capacité de lire les phrases écrites qu'il rencontre sur son chemin et d'en faire un compte rendu à la fin de son parcours. Voici un exemple du type de récit qu'on voudrait produire après un parcours du robot dans un laboratoire:

> *J'ai été dans le bureau de M. Iván V. Meza Ruiz. M. Iván V. Meza Ruiz est un ingénieur de recherche à l'IIMAS. Il travaille dans l'équipe Golem, appartenant au Département d'Informatique de l'Institut de Recherche en Mathématiques Appliquées et Systèmes. Dans son bureau, j'ai vu un affiche avec la phrase* ***Tonerre de Brest****. Cette phrase est l'expression favorite du capitaine Haddock, le célèbre personnage de Hergé dans les aventures de Tintin. Après je suis allé dans le couloir et j'ai lu sur une affiche la phrase* ***Sortie de secours****. La "Sortie de secours" est une sortie amménagée dans une pièce, un bâtiment ou un moyen de transport pour permettre une évacuation rapide des lieux par les personnes en cas de sinistre.*

Notre hypothèse est que la production de ce genre de récits est possible en embarquant le système de lecture automatique (*machine reading*) FRED (développé par l'équipe d'Aldo Gangemi, professeur à l'Université Paris 13) dans le robot Golem (développé par l'équipe de Luis A. Pineda, professeur à l'IIMAS) et en adaptant le système de génération de texte RTGen (développé par l'équipe de Claire Gardent, Directrice de recherche CNRS au LORIA à Nancy). Le robot a la capacité de se déplacer dans le labo et de reconnaître les messages textuels, tandis que FRED est capable de chercher et d'interpréter des mots clés à partir de DBPedia, la base de données de Wikipedia. Enfin l'adaptation de RTgen aux données RDF de DBPedia permettra de transformer le graphe RDF construits par FRED en un texte décrivant le parcours de FRED. Le principal critère discursif pour la génération du récit est le parcours spatial: Golem (aidé par FRED et RTGen) raconte ce qu'il a lu sur son chemin (et trouvé sur Wikipedia) .

2. Consortium

IIMAS, UNAM (Mexique)

- Luis A. Pineda C. (coordinateur)
- Iván Vladimir Meza R.

LIPN, Universidad de París 13 (France)

- Aldo Gangemi (Coordinateur)
- Jorge Garcia Flores
- Davide Buscaldi

CNRS/LORIA Nancy (France)

- Claire Gardent (Coordinatrice)
- Laura Perez-Beltrachini

3. Contexte, positionnement et objectifs scientifiques

Introduction

La robotique de service est considérée comme un axe stratégique du développement industriel dans le prochaines 20 années [1]. Un enjeu capital dans la robotique de service est la question de l'acquisition des connaissances nécessaires pour qu'un robot raisonne dans un environnement informationnel incomplet, où la structure de la tâche à accomplir ne peut pas être prévue à l'avance[2]. Certaines de ces connaissances peuvent être véhiculés par des

messages textuels présents dans l'environnement immédiat du robot. Pour qu'un robot puisse accéder au sens de ces messages, il est indispensable d'implémenter un *comportement onomasiologique*[3], c'est-à-dire un système d'algorithmes dont le résultat est un processus de lecture où le robot est capable de percevoir un message écrit dans l'environnement spatial, de transcrire ce signal visuel dans une chaîne de caractères et de déchiffrer cette chaîne pour lui attribuer une interprétation sémantique.

Le web peut être utilisé comme un amplificateur de connaissance lorsqu'un message à interpréter n'est pas présent dans la base de connaissances propre au robot [4] où lorsque le robot a besoin d'informations complémentaires par rapport à un message. Nous proposons d'étendre les capacités informationnelles d'un robot avec un système d'extraction de connaissances conçu pour le web sémantique. De cette manière on cherche à établir un lien entre la base de connaissances nécessaires pour les tâches dites *statiques* (dont le déroulement est connu à l'avance) et les tâches*dynamiques*, où le robot doit faire face à des messages et des instructions dont il peut ignorer le sens. Ces sont ces derniers qui feraient l'objet de requêtes envers le web sémantique pour retrouver une interprétation sémantique adaptée.

Comment d'ailleurs évaluer si l'interprétation choisie par le robot devant un message inconnu est correcte? Pour ce faire, nous proposons de lui attribuer aussi un *comportement sémasiologique*[3], c'est-à-dire, une capacité de générer un récit basé et sur son expérience spatiale et sur les messages qu'il a réussi à lire dans son parcours. Des algorithmes de génération de langage naturel sont donc nécessaires. La tâche que nous proposons consiste alors à donner à un robot la capacité de:

1. Effectuer un itinéraire dans l'espace (par exemple, un musée, un hôpital où l'enceinte du labo);
2. lire les messages textuels qu'il retrouve sur son parcours (par exemple, les noms de personnes sur les portes, les noms des médicaments au chevet des malades, les notices à côté des œuvres dans un musée, la signalétique dans tous les cas);
3. chercher le sens de ces messages
4. produire un compte rendu des endroits visités ainsi que de ce qu'il a lu sur son chemin.

Notre hypothèse est que cette tâche est possible en embarquant le système d'extraction de connaissances pour le web sémantique FRED[5] dans le robot Golem [2], et en y rajoutant le module de génération de langage naturel du CNRS/LORIA[6].

Dans l'état de l'art on trouve très peu de travaux où les technologies du langage sont appliqués en robotique, hors le cadre de la modélisation de dialogues. D'un point de vue recherche, la tâche qu'on propose permettrait de faire face à des verrous scientifiques interdisciplinaires. En robotique, on serait amené à modéliser les comportements nécessaires à la reconnaissance de messages textuels à partir de la caméra vidéo embarquée dans Golem, et aussi à modéliser la production orale du compte rendu à la fin de l'itinéraire avec une visée multilingue a moyen terme. En extraction de connaissances pour le web sémantique, on devra produire une version de FRED capable d'être embarqué dans un robot sans connexion Internet avec une visée également multilingue quant aux messages reçu en entrée. En génération de langage naturel, il faudra un système de génération qui permette de générer avec un minimum d'intervention humaine des description d'itinéraires de bonne qualité.

Quant à l'étendu des langues traitées, nous avons beaucoup réfléchi par rapport à la possibilité de travailler en langue française tout au long de la chaîne qui va des messages textuels repérés par le robot jusqu'au compte rendu à la fin de l'itinéraire. Étant donnée que le financement concerne des petites initiative de recherche et que les expertises complémentaires du consortium (robotique, représentation et traitement des connaissances, génération en langue naturelle) ont d'abord besoin de se ressembler sur un objectif réalisable, nous avons opté pour travailler dans un premier temps sur l'anglais, langue pour laquelle les différents modules de la chaîne ont déjà été testés, tout en ayant une visée multilingue dans le moyen terme. Si le financement AUF nous est accordé et le projet se déroule comme prévu, nous aurions surement établi les bases pour monter un projet plus ambitieux dont le but sera précisément la génération de récits en français et en espagnol. De ce fait, on s'engage à faire un effort de diffusion et vulgarisation des résultats dans des revues et des conférences francophones, ainsi qu'à créer un réseaux scientifique que dans le long terme serve de passerelle scientifique pour attirer des étudiants mexicains vers des formations doctorales chez les deux membres francophones du consortium.

Mise à part la prise en compte du multilinguisme dans le moyen terme, d'un point de vue recherche ce travail pourrait donner lieu à des collaborations intéressantes dans le domaine des modèles de cognition pour la robotique de service. Le fait d'étendre ces modèles avec des représentation de connaissances issues du web sémantique est assez innovateur et pourrait créer un bel élan scientifique entre nos équipes. De même pour le rassemblement des équipes en génération et langage naturel, dont les compétences interdisciplinaires pourraient être appliquées dans d'autres initiatives de recherche rassemblant nos laboratoires respectifs.

3.1 La robotique de service

La robotique de service consiste à étudier la structure, le comportement et les mécanismes de composition susceptibles d'être appliquées dans des tâches pratiques effectuées par une machine capable d'évoluer avec une certaine indépendance du domaine[2]. Les tâches de la compétition *Robocup@Home*[7] seraient alors des instances pratiques où des preuves de concept en environnement protégé pour des scénarios réels où le robot doit faire face à des circonstances imprévues. Ainsi, dès cette perspective, le type de tâches permet de distinguer deux types de mécanismes de composition: l'un statique, plus adapté à des tâches dont tous les données et les variations possibles sont connues à l'avance (comme dans les défis de *Robocup@Home*), et l'autre dynamique pour faire face à des scénarios où le robot doit composer avec des commandes arbitraires où des circonstances où les données de la tâche ne sont pas accessibles à priori.

Le robot Golem est un habitué du tournoi *Robocup@Home* [8][9][10]. La version actuelle du robot (Golem-III) est le résultat d'un travail entamé en 2002 avec un premier robot destiné à instancier un modèle de dialogue pour une activité conversationnelle.[11]. En 2009 une nouvelle version a été développée, avec des capacités de vision et de réaction accrues. Cette version du Golem été capable de jouer au jeu deº deviner une carte[12] avec les enfants visiteurs du musée scientifique *Universum* (UNAM). En 2010 une transformation radicale a produit Golem-II, un nouveau robot de service basé sur une architecture cognitive, un système de dialogue et un interprète de commandes, ces trois éléments étant le cœur du travail théorique. Ce à ce moment que le robot commence à participer dans la compétition *Robocup@Home*, dont il a obtenu la troisième place en 2012[9]. La version avec nous proposons de travailler est Golem-III.

Figure 1: Le robot Golem III

La tâche d'effectuer un trajet, lire les messages textuels sur le chemin et en faire un compte rendu informé à la fin de la tâche sera modélisée en s'appuyant sur le modèle de dialogue existant (structure de la tâche) où il y aurait deux comportements de base:

1. Lire les textes affichés
2. Faire un compte rendu oral du texte généré par les agents embarqués d'extraction de connaissances (FRED) et génération de langage naturel.

Il existe déjà une tâche qui consiste à [13] faire un tour dans un musée et qui pourrait servir de base à la modélisation de notre tâche, avec des interruptions déclanchées par des signaux visuels à la place des interruptions sonores de la tâche actuelle.

3.2 L'extraction de connaissances à partir des textes pour le web sémantique

L'extraction de connaissances à partir des messages textuelles ici proposée s'appuie sur les technologies du Web Sémantique, c'est à dire sur le réseau des données auto-explicites (et donc aisément lisibles par une machine[14]) disponibles sur Internet. FRED[5] est un outil de lecture pour le web sémantique conçu dans le cadre de la création d'ontologies à partir de connaissances exprimées en langage naturel, et qui s'inspire de la DRT de Kamp et de la grammaire des cas de Fillmore [15]. FRED reçoit en entrée une suite des phrases en langage naturel pour en produire un graphe en RDF (le format standard d'échange des données pour le web sémantique). L'idée est de représenter dans ce graphe un maximum de connaissances extraites à partir des textes (nom de personnes, dates, événements, sens des termes, taxonomies, relations entre entités) pour en produire un graphe orienté dont les sommets et les arêtes font référence à des données structurés disponibles publiquement, comme DBPedia (la base de données de Wikipedia) et plus généralement au nuage des données ouvertes [16].

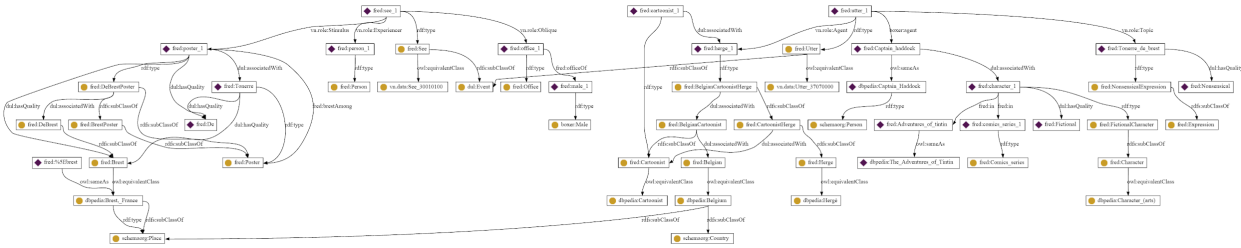

*Figure 2: Réprésentation RDF que FRED fait de la phrase: "Dans son bureau, j'ai vu un affiche avec la phrase **Tonerre de Brest**. Cette phrase est l'expression favorite du capitaine Haddock, le célèbre personnage de Hergé dans les aventures de Tintin."*

Les verrous scientifiques concernant l'extraction de connaissances textuelles avec FRED sont:

1. Le développement d'une version de FRED adapté à être embarquée dans le robot Golem, ce qui implique le téléchargement des sources ouvertes de données (comme DBPedia) pour être installées dans la mémoire locale du robot Golem, qui n'a pas de connexion Internet propre (comme c'est le cas de la plus grande partie des robots qui participe dans la coupe *RoboCup*).
2. La préparation de FRED pour le multilinguisme. Actuellement FRED n'est disponible que pour la langue anglaise. L'incorporation des ressources multilingues comme Babelnet permettrait de préparer le terrain pour la génération de récits par Golem en français et en espagnol.

3. Puisqu'il s'agit d'une *petite initiative de recherche*, le présent projet fait abstraction de certains problèmes qui se présenterait dans un cadre strictement applicatif, et notamment celui du choix parmi tous les messages textuels qu'un robot pourrait retrouver sur son chemin. Comment trier ce qui est important (comme par exemple, la signalétique faisant référence à la *sortie de secours*) de ceux qui ne mérite pas d'être rapporté dans le compte rendu (par exemple, le nom d'une marque sur une bouteille d'eau). La prise en compte du contexte phénoménologique du robot (c'est à dire, le contexte non textuel) représente un verrou scientifique qu'on pourrait prévoir à cette étape. Une solution possible à ce problème est la fouille des sous-graphes produits par FRED pour en repérer les plus saillants.
4. Le fait de considérer le web sémantique comme une source possible d'extension informationnelle pour un robot implique des ponts théoriques et technologiques entre les bases de connaissances et de raisonnement du Golem (typiquement en Prolog) et les représentation en RDF propres au web sémantique.

3.3 La génération de langage naturel

La génération de langage naturel vise à produire du texte à partir de données et d'un but communicatif (expliquer, comparer, décrire, etc.). Les données d'entrée peuvent être de différents types: données numériques, données extraites d'une base de données ou de connaissances ou même d'un texte. Avec le développement du web sémantique, les recherches sur le développement de systèmes de génération pouvant générer du texte à partir de données représentées dans le format du web sémantique (OWL, RDF, etc) se sont intensifiées. On peut distinguer trois grands types d'approches. Les approches basées sur les langages naturels controlés (*Controled Natural Language*, CNL), les approches qui apprennent un modèle statistique à partir d'un corpus parallèle alignant texte et données et les approches basées sur des patrons linguistiques.

Les approches basées sur les langages naturels controlés sont souvent des approches symboliques (c-à-d basées sur des règles) et ont plutôt été utilisées pour la verbalisation d'ontologies. Parce qu'elles font des hypothèses fortes sur la relation entre données et structures linguistiques (e.g., un axiome, une phrase et une relation, un verbe), les textes qu'elles produisent manquent souvent de naturel e.g.,

> Every cat is an animal. Every dog is an animal. Every horse is an animal. Every rabbit is an animal.

Les approches statistiques utilisent un corpus parallèle alignant texte et données pour apprendre un modèle permettant de générer du texte à partir de nouvelles données. Konstas et Lapata (2012)[17] apprennent une grammaire hors contexte probabiliste qui décrit comment (dans quel ordre et avec quels mots) les entrées et des champs de la base de données sont verbalisés dans le texte parallèle. Angeli *et al* (2010)[18] présente une approche où le processus de génération est décomposé en une séquence de décisions locales, organisées hiérarchiquement et apprises par une modèle discriminant. Wong et Moonery (2007)[19] adaptent un système de traduction automatique pour apprendre un modèle qui permet de traduire les données en texte. Si ces approches sont souvent efficaces et robustes, la qualité des phrases produites est fortement aléatoire. Un autre inconvénient majeur de ces approches est la nécessité de disposer d'un corpus parallèle de taille suffisante. Pour chaque nouveau domaine abordé, un nouveau corpus doit être créé.

Plus récemment, des approches statistiques ou hybrides symboliques/statistiques ont été proposées qui combinent patrons linguistiques et informations ou modèles statistiques pour générer du texte. Cimiano *et al.* (2013)[20] apprend la probabilité des structures syntaxiques d'un corpus spécifique au domaine considéré puis utilise ces probabilités pour sélectionner, pendant la génération, l'arbre syntaxique qui maximise un score prenant en compte la probabilité normalisée des arbres syntaxiques, les catégories syntaxiques, les synonymes et le sens lexical des mots étiquettant l'arbre et le score données pour chaque alternative par un modèle de langage. Kondadadi *et al.* (2013)[21] présentent une approche statistique dans laquelle des patrons de phrases appris automatiquement à partir de corpus textuels sont ordonnés pour chaque position dans le texte par un modèle de Machine à vecteurs de support (SVM). Enfin Perez-Beltrachini *et al.* (2013)[6] présentent une méthode hybride pour la verbalisation de requêtes sur des bases de connaissances combinant un lexique construit automatiquement à partir des noms des relations et concepts de la base de connaissances, une grammaire spécifiées manuellement et un algorithme de recherche en faisceau. Un avantage majeur, commun à ces approches, est qu'elles minimisent le besoin pour des spécifications manuelles (construction de corpus parallèles, lexiques, grammaires) tout en préservant la possibilité de générer des textes variés et de bonne qualité (pas de langage controlé).

Pour générer les descriptions d'itinéraires produites par Golem et FRED, nous nous appuierons sur un système hybride symbolique/statistique développé au LORIA par l'équipe SYNALP. Issu du modèle Quelo-RTGen présenté par Perez-Beltrachini *et al.* (2013) [6], ce modèle combine:

1. un générateur de lexique qui permet de construire un lexique automatiquement à partir des noms des relations et concepts de la base de connaissances considérés
2. une grammaire d'arbres adjoints écrites manuellemnt
3. un hypertagger statistique filtrant l'espace de recherche initial et performant la segmentation des données en sous-ensembles verbalisables chacun par une phrase.
4. un algorithme de génération

Ce système a été développé pour la verbalisation de requêtes sur des bases de connaissances et le développement d'une base de test contenant de données RDF extraites de DBPedia est actuellement en cours. Afin de l'adapter aux descriptions d'itinéraires produites par Golem et FRED, plusieurs modifications seront nécessaires:

1. extension de la grammaire pour prendre en compte les constructions spatiales, temporelles et discursives produites par Golem
2. Conception, implémentation et évaluation d'une méthode de segmentation des triplets RDF produits par FRED permettant de partitionner les données d'entrées en sous-ensembles de triplets verbalisables par une phrase. Deux pistes sont possibles, soit ré-entrainer l'*hypertagger* sur un corpus parallèle de données RDF produites par FRED et le texte correspondant; soit spécifier des règles de segmentation manuellement.
3. Adaptation du module lexical aux ressources RDF utilisées. On cherchera en particulier à utiliser les labels associés par DBPedia aux ressources RDF.
4. Développement d'un module pour le traitement des expressions référentielles pour produire des chaines référentielles appropriées. Un module minimal existe dans le système actuel. Il s'agira de l'adapter pour mieux prendre en compte le type de chaines référentielles produites par Golem+FRED

Afin d'évaluer le système résultant, il sera également nécessaire de créer une base de tests contenant des paires (données, texte) telles que les données seront des exemples de sorties produites par Golem+FRED et le texte sera une verbalisation (produite par un être humain) de

ces données. Cette base de test permettra d'évaluer les différentes composantes du système de génération (segmentation en phrase, lexicalisation, génération des expressions référentielles) et ses sorties (verbalisation des données d'entrée).

D'un point de vue recherche, le verrou principal concerne le développement d'un système de génération qui permette de générer, avec le minimum de développement manuel et dans un cadre le plus générique possible, des descriptions d'itinéraires de bonne qualité. Il s'agit d'une part, d'apprendre à partir des données du web (texte et données) des modèles permettant de lexicaliser les données, de segmenter ces données en fragments de taille verbalisable par une phrase et de choisir les expressions référentielles appropriées (*"M. Iván V, Meza Ruiz, Iván, il, Dans* son *bureau"*), et d'autre part, d'utiliser ces modèles pour produire des textes où l'interaction entre expressions spatiales (*dans son bureau, à l'IIMAS*), expressions référentielles, expressions temporelles (*après*) et marqueurs du discours (*et, pour*) est optimale.

D'un point de vue informatique, le projet permettra d'adapter un système de génération actuellement utilisé pour la verbalisation de requêtes sur des bases de connaissances, à la verbalisation de données RDF représentant des descriptions d'itinéraires.

Bien que s'appuyant sur une approche et un logiciel existant, le travail d'ingéniérie et de recherche décrit ci-dessus est conséquent. Comme ce projet est en lien direct avec le projet WebNLG (Natural Language Generation for the Semantic Web) financé par l'ANR et porté par Claire Gardent, nous projetons d'une part, de collaborer étroitement avec Laura Perez-Beltrachini post-doc sur ce projet et dévelopeuse du logiciel Quelo-RTGEn et d'autre part, de cofinancer le stage du jeune chercheur pour que celui ci soit au total de 6 mois. L'objectif est donc de recruter un étudiant sur un stage de Master Recherche pour une période totale de 6 mois.

4. Programme scientifique

Étape 1: premier parcours *in vitro*

- Début: *Septembre 2015*
- Durée: 1 mois

1: Développement des prototypes de parcours *textualisées* pour le Golem

- Durée: 0,5mois
- Responsables: JCh1, IIMAS, LIPN

2: Recherche des phrases et des mots clés trouvés par Golem avec le *web service* de FRED

- Durée: 0,25m
- Responsable: JCh1, LIPN

3: Premier parcours *in vitro* avec Fred + mots clés + génération d'un récit rudimentaire

- Durée: 0,25m
- Responsables: JCh1, IIMAS, LIPN

Étape 2: Extraction de connaissances sémantiques (*machine reading*) avec FRED et optimisation du parcours *in vitro*

- Début: *Octobre 2015*
- Durée: 6 mois

1: Embauche du jeune chercheur 1 (JCh1) par l'IIMAS

2: Développement d'une méthode d'évaluation des parcours *in vitro*

- Durée: 1m
- Responsable: JCh1, IIMAS, LIPN

3: Développement et optimisation d'un *fork* de FRED adaptée au robot Golem

- Durée: 2m
- Responsable: JCh1, LIPN

4: Développement d'une méthode de desambigüisation de mots clés dans FRED (peut être en s'appuyant sur la perception phénoménologique de Golem

- Durée: 1m
- Responsable: JCh2, IIMAS

5: Optimisation expérimentale des parcours *in vitro*

- Durée: 1m
- Responsable: JCh1

6: Documentation et publication des résultats *in vitro*

- Durée: 1m
- Responsable: JCh1

Étape 3: Génération des récits cohérents à partir des expériences spatiales de Golem

- Début: *Février 2016*
- Durée: 6 mois

1: Embauche du deuxième jeune chercheur (JCh2) par le CNRS/LORIA

2a: Méthode d'alignement statistique où de segmentation symbolique des triplets RDF produits par FRED

- Durée: 2,5m
- Responsable: JCh2, CNRS/LORIA

2b: Modification de la grammaire du système de génération du LORIA pour prendre en compte les constructions spatiales, temporelles et discursives de Golem

- Durée: 2,5m
- Responsable: JCh2, CNRS/LORIA

3a: Adaptation du module lexical aux ressources RDF utilisées

- Durée: 1,5m
- Responsable: JCh2, CNRS/LORIA

3b: Développement d'un module pour le traitement des expressions référentielles pour produire des chaines référentielles appropriées

- Durée: 1,5m
- Responsable: JCh2, CNRS/LORIA

4a: Optimisation des récits générés

- Durée: 2m
- Responsable: JCh2, CNRS/LORIA

4b: Documentation et publication des résultats

- Durée: 2m
- Responsable: JCh2, CNRS/LORIA

Étape 4: Production d'un parcours *in vivo*

- Début: *Août 2015*
- Durée: 2 mois

1: Implémentation d'un agent FRED embarqué dans le robot Golem

- Durée: 1m
- Responsable: IIMAS, CNRS/LORIA, LIPN

2: Démo publique dans le musée scientifique universitaire Universum

- Durée: 1m
- Responsables: IIMAS, LIPN

5. Rôle de la langue française

Dans un premier temps, l'objectif de ce projet est de rassembler dans un même consortium les expertises complémentaires (robotique, représentation et traitement des connaissances, génération en langue naturelle) nécessaires pour produire des textes verbalisant les observations d'un robot dans un parcours spatial et de les utiliser pour adapter et combiner des systèmes (FRED, Golem, QUELO-RTGen) existants. Le travail d'ingéniérie et de recherche nécessaire pour cette première étape étant conséquents, nous avons choisi de travailler dans un premier temps sur l'anglais, langue pour laquelle les différents modules ont déjà été testés. Néanmoins, à moyen terme nous envisageons d'utiliser les données DBPedias et les ontologies multilingues telles que Babelnet pour étendre le système proposé pour ce projet PIRAT, au français et à l'espagnol.

Nous allons diffuser les résultats de nos recherches en langue française et dans le congrès et dans la revue de l'Association pour le Traitement Automatique des Langues. Étant donné que deux tiers du consortium est francophone, la langue française sera également notre principale langue de communication. De plus, le réseaux que nous avons réuni pour ce projet à une vocation à devenir une passerelle scientifique pour attirer des étudiants mexicains vers des formations doctorales chez les deux membres francophones du consortium.

6. Stratégie de valorisation

1. Un papier en français dans la conférence TALN 2016 (Traitement Automatique des Langues)
2. Un article en français dans la revue internationale Traitement Automatique des Langues.

3. Un papier en anglais dans la conférence AAAI 2016 (Association for the Advancement of Artificial Intelligence)
4. Un papier en anglais dans une conférence sur le traitement des langues (ACL - Association for Computational Linguistics, NAACL - North American Chapter of the Association for Computational Linguistics, European Chapter of the EACL Association for Computational Linguistics, EMNLP - Empirical Method for Natural Language Processing ou INLG - International Conference on Natural Language Generation)

Les parties fixeront les modalités d'exécution du projet et les règles de dévolution et d'exploitation des droits de propriété intellectuelle des résultats du projet dans le cadre d'un accord de consortium.

7. Perspectives

Du point de vue de la robotique, ce projet permettrait d'explorer les liens entre la base de connaissances de Golem[22] en Prolog et des extensions dynamiques possibles envers le web sémantique et le format RDF. L'idée de que devant des situations imprévues le robot puisse s'appuyer sur le web sémantique pourrait donner lieu des nouvelles méthodes pour structurer des comportements dynamiques dans les modèles cognitifs en robotique. De plus, ce projet permettrait également de préparer le terrain pour une prise en compte du multilinguisme à moyen terme, et en particulier pour la lecture et production de récits en langue française.

Le multilinguisme viendrait également enrichir FRED[5], l'outil d'extraction de connaissances pour le web sémantique, qui actuellement n'a été testée qu'en langue anglaise. Bien que dans ce projet la seule langue traité sera l'anglais, le financement de notre projet par l'AUF renforcerait la perspective de production de récits en français et en espagnol à moyen terme. Par ailleurs, le fait d'embarquer FRED dans des dispositifs sans connectivité Internet pourrait donner lieu à une nouvelle famille d'applications embarquées issues du web sémantique.

Du point de vue de la génération de texte, ce projet permet de mettre en place une architecture pour la génération de textes à partir de données RDF. Comme l'indique [23], pour les être humains, les standards (e.g., RDF, OWL) établis par la communauté du web sémantique pour représenter les données et les ontologies sont difficiles à comprendre et à manipuler. Avec le développement du web sémantique, la croissance rapide des données liées (linked data), la prolifération des bases de connaissances et plus généralement, avec l'émergence des données massives, il y a un besoin accru de développer des technologies qui permettent aux être

humains un accès simple et naturel aux données orientées machine du web des données. Parce qu'il permet de convertir les données en texte, le système de génération développé dans le cadre de ce projet PIRAT, procure un moyen naturel de présenter ces données de façon intuitive, structurée et cohérente. Plus généralement, de tels systèmes permettent d'expliciter le contenu de données liées ou de bases de connaissances à des utilisateurs non expert; de générer des explications, des descriptions et des résumés à partir de DBPedia ou d'autres bases de connaissances; de guider l'utilisateur dans la formulation de requêtes sur des bases de connaissances; et de présenter l'information contenue dans les données liées publiées par les institutions pour l'héritage culturel telles que les musées et les bibliothèques (cf les exemples listés ici), sous différentes formes (par exemple, pour un utilisateur expert ou novice) et dans différentes langues. De fait, la génération est de plus en plus vue comme "changeant les règles du jeu" (game changing) et de nouvelles compagnies ont récemment émergé qui vise à "faire communiquer les données massives directement, pas en nombres ou en tables qui exigent analyse et explication, mais dans des textes narratifs riches dont on pourrait penser qu'ils ont été produits par un expert humain" (e.g., ARIA).

8. Références

1. Aurélie Sobocinski. *Quel avenir pour la robotique de service ?* CNRS Le journal, 28.07.2014.
2. Luis A. Pineda, Arturo Rodríguez, Gibran Fuentes, Caleb Rascon and Ivan V. Meza. *Concept and Functional Structure of a Service Robot.* International Journal of Advanced Robot Systems, 2015, 12:6.
3. Bernard Pottier. Sémantique générale. *Presses universitaires de France*, Paris, 1992.
4. Kai Zhou, Michael Zillich, Hendrik Zender and Markus Vincze. *Web Mining Driven Object Locality Knowledge Acquisition for Efficient Robot Behavior.* 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 7-12, 2012. Vilamoura, Algarve, Portugal.
5. Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero. *Frame-based detection of opinion holders and topics: a model and a tool.* IEEE Computational Intelligence, 9(1), 2014
6. Laura Perez-Beltrachini, Claire Gardent and Enrico Franconi. *Incremental Query Generation*. EACL 2014,. Gothenburg, Sweden, April 2014.

7. Sven Wachsmuth, Dirk Holz, Maja Rudinac, Javier Ruiz-del-Solar. *RoboCup@Home - Benchmarking Domestic Service Robots.* Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas.
8. Luis Pineda, *The Golem Group: The Golem Team*, *RoboCup@Home 2011*. Proceedings of Robocup 2011. vol --, pp 8. 2011.
9. Luis Pineda, *The Golem Group: The Golem Team, RoboCup@Home 2012*. Proceedings of Robocup 2012. vol , pp . 2012.
10. Luis A. Pineda, *Grupo Golem: RoboCup@Home 2013*. Proceedings of Robocup 2013. vol , pp . 2013.
11. Edith Moya, E., Hernández, M., Pineda, L. and Meza, I.: *Speech Recognition with Limited Resources for Children and Adult Speakers.* Tenth Mexican International Conference on Artificial Intelligence - Special Session - Revised Papers. vol 2276, pp 57-65. 2011.
12. Ivan Meza, Salinas, L., Pavón, E., Avilés, H. and Pineda, L.: *A Multimodal Dialogue System for Playing the Game "Guess the card".* Procesamiento de Lenguaje Natural. . vol 44, pp 131-138. 2010.
13. Caleb Rascon, Ivan Meza, Gibran Fuentes, Lisset Salinas and Luis A. Pineda. *Integration of the Multi-DOA Estimation Functionality to Human-Robot Interaction.* Int J Adv Robot Syst, 2015, 12:8
14. Oren Etzioni , Michele Banko , Michael J. Cafarella, *Machine reading*. Proceedings of the 21st national conference on Artificial intelligence, p.1517-1519, July 16-20, 2006, Boston, Massachusetts
15. V. Presutti, F. Draicchio and A. Gangemi, *Knowledge extraction based on discourse representation theory and linguistic frames.* EKAW2012 Conference, LNCS, Springer, 2012.
16. Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak, *Linking Open Data cloud diagram 2014*, http://lod-cloud.net/
17. Konstas, Ioannis, and Mirella Lapata. *Unsupervised concept-to-text generation with hypergraphs*. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 3 Jun. 2012: 752-761.

18. Angeli, Gabor, Percy Liang, and Dan Klein. *A simple domain-independent probabilistic approach to generation.* Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing 9 Oct. 2010: 502-512.
19. Wong, Yuk Wah, and Raymond J Mooney. *Generation by Inverting a Semantic Parser that Uses Statistical Machine Translation.* HLT-NAACL 2007: 172-179.
20. Cimiano, Philipp et al. "Exploiting ontology lexica for generating natural language texts from RDF data." (2013).
21. Kondadadi, Ravi, Blake Howald, and Frank Schilder. *A Statistical NLG Framework for Aggregated Planning and Realization.* ACL (1) 6 Aug. 2013: 1406-1415.
22. Luis A. Pineda, Lisset Salinas, Ivan V. Meza, Caleb Rascon and Gibran Fuentes. *SitLog: A Programming Language for Service Robot Tasks.* Int J Adv Robot Syst, 2013
23. A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, and C. Wroe. *Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns.* Engineering Knowledge in the Age of the Semantic Web, pages 63–81, 2004.

Budget

Budget global estimé: € 15000

Dont subvention demandée à l'AUF: € 12000

Concept	**Durée**	**Montant**
Mission d'Aldo Gangemi (LIPN) à Mexico (IIMAS)	5 jours	€2000
Mission de Luis Pineda (IIMAS) à Villetaneuse (LIPN)	5 jours	€2000
Mission de Claire Gardent à l'IIMAS (Mexico)	5 jours	€2000
Embauche d'un jeune chercheur par l'IIMAS pour le développement de l'interface FRED+Golem en el IIMAS	720 heures sur 6 mois	€3000
Embauche d'un jeune chercheur par le CNRS/LORIA pour la génération de récits cohérents à partir des expériences spatiales du robot Golem (CNRS/LORIA complémentera ce budget à hauteur de 3000 euros pour permettre une embauche sur 6 mois)	720 heures sur 6 mois	€6000
		€15000

Échéancier des versements

Versement à l'UNAM Date: 1er septembre 2015	**€5000**
Versement à l'Université Paris 13 Date: 4 janvier 2016	**€2000**
Versement au CNRS/LORIA Date: 4 février 2016	**€5000**
Total versé par l'AUF:	**€12000**