Vers une nouvelle structuration de l'information extraite automatiquement

Alejandro ACOSTA
LATTICE-CNRS (UMR 8094)
Université Paris 7
alejandro.acosta@linguist.jussieu.fr

Résumé. Les systèmes d'Extraction d'Information se contentent, le plus souvent, d'enrichir des bases de données plates avec les informations qu'ils extraient. Nous décrivons dans cet article un travail en cours sur l'utilisation de données extraites automatiquement pour la construction d'une structure de représentation plus complexe. Cette structure modélise un réseau social composé de relations entre les entités d'un corpus de biographies.

Abstract. Information Extraction systems are widely used to create flat databases of templates filled with the data they extract from text. In this article we describe an ongoing research project that focuses on the use of automatically extracted data to create a more complex representation structure. This structure is a model of the social network underlying the relations that can be established between the entities of a corpus of biographies.

Mots-clés : extraction d'information, analyse de réseaux sociaux, biographies, entités nommées, représentation de connaissances.

Keywords: information extraction, social network analysis, named entities, biographies, knowledge representation.

1 Introduction

Les systèmes de Traitement Automatique de Langues (TAL) sont nés, en grande partie, du désir de modéliser la compréhension du langage naturel et de créer des systèmes capables de transformer un discours incompréhensible pour une machine en une représentation formelle explicite de l'information véhiculée par le langage. Cette tâche s'est avérée plus difficile que prévu et en conséquence les traducteurs automatiques, ainsi que les systèmes intelligents autonomes restent du domaine de la science-fiction.

Le domaine de la Représentation de Connaissances s'est concentré sur le problème de la structuration des contenus, et sur les opérations sur les structures de représentation. De son côté, la Compréhension du Discours s'est heurtée à la complexité de l'interprétation du langage naturel et a dû chercher à être plus modeste dans ses objectifs.

L'Extraction d'Information (EI) est née du désir de construire des systèmes capables de répondre à des tâches spécifiques de compréhension. Son évolution est fortement liée aux campagnes d'évaluation de systèmes capables de repérer des événements ponctuels, ainsi que les

acteurs associés à ces événements. Les évaluations ont encouragé les avancées dans le domaine, mais les techniques d'évaluation ont aussi influencé la définition d'un domaine qui, à l'origine, était celui de la compréhension des textes. Cette influence a dirigé l'EI vers des applications où l'on cherche à repérer des faits spécifiques dans les textes d'un domaine particulier. Ces systèmes sont capables de collecter un ensemble d'informations, structurées dans des formulaires qui remplissent des bases de données plates.

Ces bases de données peuvent être utilisées facilement pour l'évaluation des systèmes en termes de rappel et de précision (et de leurs mesures dérivées). Or, il est plus difficile d'évaluer des structures plus complexes, mettant en rapport les événements entre eux, ou les acteurs qui se répètent dans les différents événements (une représentation plus proche des bases de données relationnelles qui sont devenues le standard). De ce fait, on s'est contenté de construire des systèmes capables de faire ce qui était nécessaire pour leur évaluation et qui n'ont pas cherché à construire des structures de représentation plus riches.

Nous pensons que les techniques d'EI peuvent être utilisées pour construire des représentations plus complexes, des structures relationnelles. La volonté de structurer l'information dans des formulaire isolées a joué un rôle dans le choix des méthodes qui permettaient d'atteindre ce but. De la même manière, l'objectif de constituer une représentation plus complexe va aussi avoir une influence sur la définition de la tâche d'extraction.

Nous présentons dans cet article un travail en cours qui propose d'utiliser autrement les résultats d'une tâche d'extraction. Plus précisément, nous proposons une structure de représentation plus riche que celle qui est suggérée dans les tâches conventionnelles. Pour ce faire, nous présentons un travail mené sur un corpus qui se prête bien à la conception d'une représentation structurée : une vaste collection de notices biographiques.

Le reste de ce document est organisé de la manière suivante : dans la section 2 nous présentons le corpus de biographies, dans la section 3 nous discutons les motivations pour chercher une représentation plus complexe de l'information extraite, nous décrivons les éléments qui composent cette représentation (3.1 et 3.2) et nous parlons de l'état de l'art (3.3). Dans la section 4 nous présentons les techniques d'EI utilisées. Finalement, nous présentons quelques perspectives pour la continuation de ce travail dans la section 5.

2 Le Maitron

Le corpus utilisé dans ce travail est une collection de notices biographiques. Ces biographies peuvent être étudiées séparément mais il est aussi intéressant d'étudier l'objet plus large, et plus complexe, dont elles font partie.

Le *Dictionnaire biographique du mouvement ouvrier francais*[1] (dorénavant *Le Maitron*) est un dictionnaire contenant des notices biographiques sur les vies de milliers de personnes ayant participé activement aux luttes sociales de l'histoire française depuis 1789. Le dictionnaire original est divisé en quatre sections, correspondant à quatre périodes historiques qui vont de la Révolution Française au début de la Deuxième Guerre mondiale. La publication a commencé à s'enrichir d'une nouvelle section en 2006 avec l'ajout d'une nouvelle période entre la Deuxième Guerre et 1968.

[1] `http://www.maitron.org`

Plus de 600 auteurs ont participé au projet depuis sa création par Jean Maitron dans les années cinquante. Les notices biographiques des 56 volumes du dictionnaire (plus de 100 000 articles) sont constamment révisées et retravaillées par une équipe d'historiens dirigée par Claude Pennetier au Centre d'histoire sociale du XX^{e} siècle[2]. Le corpus du *Maitron* compte ainsi plus de 18 millions de mots, couvrant une période qui s'étend sur plus de deux siècles d'histoire.

Notre objectif est de proposer une représentation qui rende compte de la complexité du contenu du corpus (sans qu'elle soit exhaustive pour autant), et qui permette d'exploiter ces informations sous une autre forme.

3 Structuration des informations extraites

Les différentes approches traitant de la représentation de connaissances ont proposé un grand nombre de modèles dans l'étude des structures qui peuvent représenter symboliquement des informations complexes. Par ailleurs ces approches ont également proposé des techniques visant l'utilisation de ces structures pour le raisonnement automatique. Cependant, on est toujours incapable de traduire automatiquement le langage naturel en une représentation des connaissances qu'il transmet. On est malheureusement encore loin d'atteindre le niveau de performance que l'on espérait dans le domaine de la compréhension des textes il y a quelques décennies.

Le domaine de l'EI, de son côté, s'est contenté de la tâche qui consiste à récolter l'information dans des bases de données (Pazienza, 1999). Les *Message Understanding Conferences* de la fin des années 80 et des années 90 ont vu la transformation progressive de cette compréhension de messages en remplissage de descriptions d'événements (Poibeau, 2003).

l'EI est née pour affronter les difficultés rencontrées par les systèmes de compréhension des textes et chercher un compromis entre les limitations des systèmes et la qualité des résultats que l'on pouvait obtenir. La compréhension s'est vue réinterprétée et divisée en tâches successives, dont le but ultime était le remplissage de formulaires. Dans cette optique, chaque formulaire rempli correspond à une assertion sur le monde (on extrait l'information qu'on suppose vraie) ; l'ensemble de ces assertions constitue une base de données d'informations sur le domaine. Les systèmes d'EI se sont depuis penchés sur l'exploitation et le traitement d'importantes quantités de textes désormais disponibles en format numérique (et qui peuvent donc facilement être traités automatiquement).

Nous pensons que le compromis entre les limites des techniques et la complexité de la représentation peut encore être négocié. Certes, l'interprétation automatique du langage naturel, qui nous permettrait de constituer des bases de connaissances expressives et robustes, est encore hors de notre portée. Nous pensons cependant que les résultats d'un processus d'EI pourraient être améliorés en créant une sortie structurée de manière plus riche que la seule énumération des informations dans une base de données.

A notre avis, la structuration des informations pourrait être conçue autour des relations entre entités (dans le sens, grosso modo, d'entités nommées). Nous cherchons à établir des relations entre entités, plutôt qu'à repérer des événements pour remplir des formulaires avec des détails sur ces événements.

Le modèle de représentation que nous avons choisi pour traiter les informations extraites du

[2]CNRS / Université Paris I

Maitron s'inspire d'une technique de modélisation issue des sciences humaines : l'analyse des réseaux sociaux.

L'Analyse de Réseaux Sociaux (ARS) est une technique utilisée dans plusieurs domaines de recherche en sciences humaines pour étudier les relations entre individus. Ces individus sont souvent en rapport du fait de leur appartenance à un groupe ou une organisation. L'ARS s'est développée avec l'étude des rapports et interactions entre les individus de communautés particulières comme les habitants d'une île, les employés d'une entreprise ou les membres d'une bande d'adolescents d'un quartier, pour citer quelques exemples.

Avec l'avènement de l'ère de l'information et des communautés dites virtuelles de l'Internet, l'ARS attire l'attention d'une communauté scientifique de plus en plus large. Grâce à cette évolution du domaine, les chercheurs en sciences humaines, qui ont l'habitude de mener leurs recherches sur des corpus de taille limitée, commencent à profiter des exploits d'autres disciplines qui leur permettent d'étudier les relations entre individus à un autre niveau.

La théorie des graphes, pour citer un exemple, a développé des méthodes de représentation de réseaux et des techniques d'exploration qui sont maintenant utilisées pour interpréter les données relationnelles.

Le remplissage de formulaires avec des données biographiques n'est pas, en soi, une tâche qui permette d'enrichir l'information qui est déjà organisée dans notre corpus. En revanche, nous pouvons construire une structure à partir des nombreux liens entre les personnages du *Maitron*, une structure qui pourra être exploitée par les chercheurs qui utilisent le dictionnaire comme ressource de recherche.

Les sections 3.1 et 3.2 sont consacrées à la description des structures utilisées pour modéliser l'information contenue dans le corpus du *Maitron*. Nous considérons qu'une modélisation inspirée de l'ARS nous permet d'atteindre un niveau de structuration plus riche que celui inspiré du remplissage de formulaires d'extraction.

3.1 Le réseau social du *Maitron*

Si l'on voit le *Maitron* non pas comme une collection de notices biographiques mais comme un corps structuré d'informations à propos d'un ensemble d'individus, sa structuration sous la forme d'un réseau social apparaît comme une organisation presque naturelle pour ces données. En effet, le Maitron est composé d'histoires de vie d'individus qui étaient souvent en rapport les uns avec les autres, et c'est sur ces rapports que s'est bâtie une partie importante de l'histoire sociale de France. On remarquera que l'on peut associer ces individus à d'autres « acteurs de l'histoire » (publications, partis politiques, etc.). Les vies de ces personnes se rencontrent à l'intérieur d'événements spécifiques, à travers leur appartenance à des groupes ou à des associations. Les liens, une fois explicités, deviennent une abstraction structurée de l'interaction entre les objets du réseau.

Une structuration de ce type imite aussi une partie de la démarche de l'historien chercheur qui se sert de ce type de ressource documentaire : la première tâche du chercheur consistant à suivre les liens entre individus, associations, événements et autres entités saillantes. Individuellement, des données biographiques plus traditionnelles comme les dates de naissance, les lieux de naissances et autres, sont toujours des attributs porteurs de sens mais secondaires dans la structuration du corpus comme un tout.

On peut orienter le processus d'extraction vers la création d'une représentation qui aurait la structure d'un réseau. Le réseau peut être modélisé comme un graphe et la tâche d'extraction consiste à récolter les données nécessaires pour construire cette représentation (les liens entre les sommets du graphe). Tout comme le résultat d'une tâche classique d'EI le lien entre deux sommets est porteur de sens en tant qu'assertion d'un fait de l'univers (l'univers du corpus). En revanche, le sens du lien est plus riche car, dans le réseau, le même lien peut être interprété aussi en fonction de son appartenance à une structure plus complexe.

Sachant donc que nous cherchons à construire une représentation sous la forme d'un réseau, il nous reste à décrire les objets qui le composent : les sommets et les arcs de cette structure.

3.2 L'ontologie du *Maitron*

Nous avons décidé de structurer les objets qui composent notre représentation, le réseau, dans une ontologie. Très utilisées dans de nombreuses applications visant le passage du Web actuel au Web Sémantique, les ontologies permettent de modéliser les connaissances d'un domaine et de spécifier son vocabulaire de représentation (Gruber, 1993). Dans notre cas, le fait d'associer des éléments qui composent le réseau à une ontologie nous permet d'organiser les objets qui nous semblent pertinents pour la construction du réseau du *Maitron*. L'ontologie est une organisation qui se prête aussi à l'évolution de notre compréhension de ces objets car elle nous permet de prévoir des niveaux de granularité. Finalement, cette manière de décrire les objets qui composent notre représentation est au cœur de sa portabilité (voir section 5).

Dans un premier temps, nous nous intéressons principalement au niveau le plus générique d'interaction entre les objets du réseau. Le point de départ, inspiré de l'ARS, sont les individus et leur relations. Dans le graphe qui représente le réseau, les individus correspondent donc aux sommets, et les relations aux arcs. Nous considérons cependant que le réseau du *Maitron* est structuré en bonne partie par l'interaction entre individus et d'autres types d'entités.

Au premier niveau de notre ontologie, nous plaçons donc deux types d'objets, les RELATIONS, et les ENTITÉS. Nous distinguons quatre types d'entités : INDIVIDUS, ORGANISATIONS, PUBLICATIONS et ÉVÉNEMENTS. Quant aux relations, elles sont définies en fonction des différents rapports qu'entretiennent deux entités. Chaque entité est susceptible d'être en relation avec une autre : les individus peuvent être en relation avec d'autres individus, mais aussi avec des organisations, des publications ou des événements. Dans la figure 1 nous trouvons la hiérarchie d'objets (à gauche) et la combinatoire des relations (à droite).

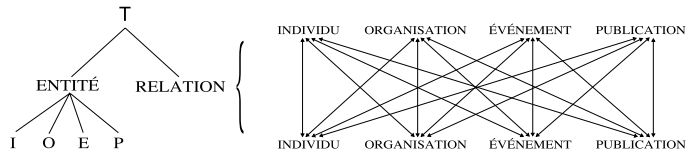

FIG. 1 – Ontologie des objets et combinatoire de relations entre entités

En langue naturel ces connections peuvent prendre des formes différentes. Par exemple, si l'on pense aux liens possibles entre INDIVIDUS et les quatre autres types d'ENTITÉS, on trouvera rapidement des marqueurs (des éléments de la langue) qui expriment ces relations : ainsi, on dira

que les individus *rencontrent* des individus, *participent* à des événements, sont *membres* d'organisations, *écrivent* pour des publications. Pour chaque type de relation il y a de nombreuses manières d'exprimer les liens entre les deux types d'entités concernés.

Pour construire le réseau de ces relations, les structures linguistiques qui sont utilisées pour les exprimer doivent être identifiées, modélisées et enfin trouvées dans le corpus. En retrouvant les instances spécifiques on peut extraire les informations nécessaires pour décrire les liens. Nous utilisons des techniques d'EI sur du texte prétraité à l'aide d'outils de TAL pour faire cette extraction.

3.3 L'extraction d'entités et de relations

D'autres travaux d'EI se sont déjà intéressés à l'extraction automatique de relations et à la représentation d'informations biographiques extraites de manière automatique. Nous présenterons brièvement quelques uns.

Riloff (1996) propose un système capable de générer ce qu'il appelle des *case frames* (qui seraient grosso modo équivalents aux formulaires d'extraction) comportant plus de deux entités dans un scénario d'extraction. Le repérage des relations entre entités est utilisé dans ce système dans le but de construire automatiquement des formulaires d'extraction plus complexes. L'objectif de Riloff est de proposer un système qui s'adapte facilement à un nouveau domaine (et une nouvelle tâche) d'extraction. L'extraction même consiste, encore une fois, à retrouver l'information nécessaire pour remplir des formulaires isolés.

Le point de départ du système REES (Aone & Ramos-Santacruz, 2000) est deux ontologies. Une première ontologie est composée d'une typologie d'entités et les relations qu'on leur associe, la deuxième ontologie est une classification de types d'événements. Les relations sont définies entre les quatre types d'entités génériques et les attributs qu'on leur associe généralement. Pour un type d'entité comme « personne », par exemple, on définit des relations avec un titre, une nationalité, un numéro de téléphone, une affiliation, un type, un soustype, etc. Quant aux événements, ils sont composés d'un ensemble de participants (les entités). L'approche cherche à donner de la généricité au système d'EI par une définition *a priori* des événements possibles ; des événements qui sont présentés comme plus génériques. Il n'y a pas de liens entre les instances isolées d'événements qui sont extraites.

Le Priol (2001) utilise la méthode dite d'exploration contextuelle pour extraire automatiquement des relations sémantiques comme, par exemple, *est partie de* ou est *inclu*. A la différence des systèmes d'extraction classiques, le système de Le Priol se concentre sur les liens sémantiques entre unités lexicales, et ne s'intéresse pas aux événements. Du fait de mettre des unités lexicales en relation, le système vise, comme nous, de produire une représentation complexe. En revanche, cette représentation est plus proche d'une ontologie extraite automatiquement pour représenter des connaissances génériques que d'un réseau social qui représente des informations liées à des événements et des acteurs.

Shinyama and Sekine (2006) utilisent des techniques de *clustering* statistique pour découvrir automatiquement des relations dans un corpus de dépêches de presse. Leur objectif est de constituer l'ensemble de relations qui peuvent être utilisées plus tard pour construire des formulaires d'extraction. Les formulaires résultent des régularités dans les structures linguistiques du corpus. Cette approche pourrait être utilisée dans le cadre de notre travail. En effet, les ingénieurs linguistes qui développent des patrons d'extraction pourraient utiliser le *clustering* statistique

pour découvrir des structures récurrentes.

Finalement, Kevers (2006) s'intéresse à l'extraction automatique de données biographiques et à leur représentation. Son objectif est de décrire un modèle de représentation d'information biographique fondé sur des triplets (*sujet, relation, objet*). Son modèle s'applique donc à la description des attributs biographiques que l'on a mentionnés dans la section 3.1 comme étant secondaires à la structuration du réseau de relations entre entités. Son travail peut donc être vu comme étant complémentaire au nôtre.

4 Extraction de structures linguistiques

Maintenant que nous avons une meilleure compréhension de la structure de la représentation que l'on veut produire et des objets qui la composent, nous devons décrire les techniques utilisées pour récolter et arranger ces objets dans cette structure.

On distinguera deux étapes dans ce processus. La première consiste à analyser le texte non structuré, c'est-à-dire à l'enrichir avec des annotations linguistiques. La deuxième étape est l'extraction elle-même, qui opère sur les structures qui résultent de la première étape.

4.1 Annotation linguistique du corpus

Nous utilisons des outils de l'architecture de développement linguistique MACAON[3] pour le prétraitement des textes du corpus. Ce prétraitement est indépendant de l'extraction et pourrait se faire avec des outils différents. C'est pour cette raison que nous n'entrons pas ici dans les détails techniques de ces outils de TAL, et nous nous limitons à la description des caractéristiques pertinentes pour la suite.

Les outils MACAON sont une collection de modules de TAL dévéloppés pour des tâches spécifiques d'annotation de textes. Avant d'utiliser les patrons d'extraction associés aux différents types de relations entre entités, le texte des biographies est analysé avec des modules MACAON qui s'occupent des tâches suivantes :

1. Segmentation en phrases
2. Tokenisation
3. Reconnaissance d'entités nommées
4. Analyse lexicale
5. Étiquettage morpho-syntaxique
6. Analyse morphologique
7. Analyse syntaxique partielle

La reconnaissance d'entités nommées consiste à repérer des formes superficielles d'objets de type ENTITÉ de l'ontologie présentée dans la section 3.2.

L'analyseur partiel suit le modèle proposé par Abney (1996) mais incorpore aussi le concept de tête. Dans le modèle d'analyse partiel de Abney, des grammaires régulières décrivant des

[3] `http://code.google.com/p/macaon/`

constituants partiels (ou *chunks*) s'appliquent en cascade aux séquences de parties du discours qui composent les phrases du texte à analyser. Dans l'implémentation de MACAON, les patrons réguliers qui décrivent les *chunks* peuvent identifier l'un de ses composants comme étant la tête de l'ensemble. Cette application, inspirée de certains formalismes linguistiques (eg. HPSG), permet au module d'extraction d'accéder plus aisément à certains éléments de la phrase.

Après le passage par cet ensemble de modules de TAL, les phrases du corpus sont interprétées[4], au moment de l'extraction, comme des séquences de structures de traits. Les structures de traits contiennent l'ensemble des annotations linguistiques ajoutées par les différents modules.

4.2 FSMs et structures de traits

La reconnaissance des fragments de texte associés aux liens qui composent le réseau, ainsi que l'extraction des arguments dans chaque instance d'une relation sont tous les deux accomplis avec des machines à états finis (FSMs) à base de structures de traits (ou FS-FSMs[5]).

Nous venons d'expliquer que la version analysée des phrases du corpus est composée de structures de traits. Le module d'extraction utilise des machines à états finis pour reconnaître des patrons non pas sur un alphabet de symboles mais sur un alphabet de structures de traits. A la différence des automates traditionnels de reconnaissance, les FS-FSMs n'utilisent pas la relation de l'égalité lors de la reconnaissance. Cette relation entre les symboles de l'entrée de l'automate est remplacée par la relation de subsomption entre structures de traits.

Dans une machine à états finis habituelle, la reconnaissance se fait au fur et à mesure que la machine change d'état, et ceci en fonction des symboles qu'elle trouve sur la bande d'entrée et des symboles associés à ses transitions. Pour changer d'état, une FS-FSM qui se trouve dans un état donné doit avoir une transition associée à une structure de traits qui subsume la structure de traits de la bande d'entrée. Parallèlement, les entités sont extraites durant la reconnaissance. Ce n'est que lorsque la FS-FSM atteint un état d'acceptation (après avoir traversé un certain nombre de transitions) que l'instance de relation est reconnue et que l'information extraite est validée. Les instances incomplètes sont rejetées.

Une machine d'extraction simple décrit un patron spécifique utilisé pour la reconnaissance d'une relation. Ce patron pourrait être, par exemple, le patron décrivant des phrases avec le nom d'une personne (une entité de type INDIVIDU) suivi de « était membre de », suivi à son tour d'une entité nommée reconnue comme le nom d'un parti politique (une entité de type ORGANISATION). Des machines plus complexes peuvent être définies pour des patrons plus expressifs ; par exemple, pour l'ensemble de patrons construits autour du déclencheur lexical *membre*.

L'utilisation d'une technique de traitement derivée des machines à états finis nous permet de tenir compte de leurs propriétés et d'utiliser les opérations définies sur elles en tant que modèle de calcul. Nous pensons plus précisément à l'union, la minimisation et la déterminisation. Le module d'extraction peut être vu comme l'union des FS-FSMs associées aux relations entre les différents types d'entités. La minimisation et la déterminisation pourraient[6] être utilisées pour

[4]Voir (Gazdar *et al.*, 1988) pour une description de l'utilisation des structures de traits pour représenter des catégories syntaxiques.

[5]On préfère l'acronyme de la version anglaise « Feature Structure Finite State Machines » du fait de l'usage courant dans la littérature de l'acronyme FSM pour les machines à états finis

[6]Nous n'avons pas encore formalisé l'application des stratégies de minimisation ni de déterminisation à des

améliorer la performance (en temps et en espace) du module d'extraction.

Par ailleurs, le processus de la construction de la représentation peut aussi être vu comme un processus incrémental. Nous n'avons pas besoin de définir toutes les relations (et tous les patrons associés à ces relations) au préalable et la totalité du réseau ne doit pas être la sortie d'une seule extraction. L'extraction, telle qu'elle est décrite ici, peut opérer comme une méthode permettant l'acquisition incrémentale de connaissances. La taille de l'ensemble de relations augmente, ainsi que la taille du réseau, au fur et à mesure que les patrons des FS-FSMs sont affinés.

A l'heure actuelle, nous nous concentrons sur la mise au point des outils décrits dans la section 4.1 et sur la modélisations des patrons d'extraction, ce qui nous permettra de passer à l'étape d'évaluation des résultats. Bien que l'on ne puisse pas présenter pour l'instant des résultats concrets, nous pouvons mentionner les perspectives qui guident notre travail.

5 Perspectives

Nous voyons, au moins, trois directions d'évolution du travail que nous menons sur le *Maitron* : l'étude du réseau qui résulte de l'extraction, l'adaptation de la structure de représentation à d'autres domaines d'application et l'inclusion de connaissances extraites dans des systèmes intelligents.

Nous avons mentionné que l'ARS attire l'attention de plusieurs domaines de recherche. On pourrait adapter les avancées dans d'autres domaines à notre utilisation du corpus. L'application de méthodes de théorie des graphes, par exemple, pourrait aider à exploiter la structure du réseau. Avec les méthodes d'analyse de cette théorie, on peut envisager de retrouver les entités les plus saillantes, ou celles qui créent des ponts entre différents fragments du réseau. Nous pouvons aussi exploiter la structure du réseau pour raffiner la classification des entités du domaine dans le but, par exemple, d'adapter l'exploitation du réseau à la construction automatique d'ontologies.

Nous nous intéressons dans notre travail aux relations entre différents types d'entités. Bien que ces relations semblent spécifiques à l'interprétation de la structure du corpus du *Maitron*, elle ne sont pas limitées au domaine des corpus biographiques. Tant que les mêmes types d'entités sont concernées, les mêmes relations et les mêmes patrons peuvent être utilisés pour des tâches d'extraction et d'acquisition de connaissances dans d'autres domaines. L'Internet et ses communautés virtuelles, par exemple, sont un vaste domaine dans lequel on trouve les mêmes types d'entités. Les textes (ou autres données) partagés par les membres de ces groupes peuvent être utilisés pour constituer un corpus qui permettrait de recréer le réseau de leurs relations.

Nous avons choisi de représenter le réseau avec des déclarations RDF[7]. Ces dernières peuvent être intégrées directement dans les systèmes experts qui sont élaborés suivant les standards XML proposés pour la représentation et l'exploitation des données en ligne. De plus, ces standards entretiennent un lien étroit avec le domaine de la représentation de connaissances ; notons que notre modélisation du réseau peut être traduite en termes de logique de description. Le réseau du *Maitron* structure des connaissances très précises, mais ces connaissances peuvent s'intégrer dans des systèmes qui exploitent des bases de connaissances pour d'autres domaines

machines qui utilisent la subsomption comme relation de reconnaissance. Leur union, en revanche, est triviale.

[7]La présentation de ces standard XML étant au-délà de la portée de cette article, nous dirigeons le lecteur vers la documentation qui peut se trouver en ligne.

d'application comme, par exemple, les systèmes de question-réponse.

6 Conclusion

Nous avons proposé d'utiliser des techniques d'EI dans le but de construire des structures de représentation plus complexes que les bases de données plates qui sont souvent utilisées pour l'évaluation des systèmes d'extraction.

Nous avons décrit une représentation organisée autour des relations entre types d'entités génériques. Le cas du travail en cours avec un corpus de biographies a été utilisé pour décrire une application concrète de cette manière d'exploiter les résultats d'un processus d'extraction dans le but de constituer une base de connaissances avec une structure inspirée de l'ARS.

Nous avons décrit la méthodologie et les outils utilisés dans le processus d'extraction, et nous nous sommes particulièrement intéressé à la description des machines à états finis qui modélisent des structures du langage naturel.

Finalement, nous avons présenté trois chemins distincts d'évolution de ce projet de recherche.

Références

ABNEY S. (1996). Partial Parsing via Finite-State Cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, p. 8–15, Prague.

AONE C. & RAMOS-SANTACRUZ M. (2000). REES: A Large-scale Relation and Event Extraction System. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

GAZDAR G., PULLUM G. K., CARPENTER R., KLEIN E., HUKARI T. E. & LEVINE R. D. (1988). Category Structures. *Computational Linguistics*, **14**(1), 1–19.

GRUBER T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. GUARINO & R. POLI, Eds., *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands: Kluwer Academic Publishers.

KEVERS L. (2006). L'information biographique: modélisation, organisation et extraction en base de connaissances. In *Actes de TALN-RECITAL*, p. 680–689, Leuven, Swisse.

LE PRIOL F. (2001). Identification, interprétation et représentation de relations sémantiques entre concepts. In *Actes de TALN*.

M. T. PAZIENZA, Ed. (1999). *Information Extraction: Towards Scalable, Adaptable Systems*, volume 1714 of *Lecture Notes in Artificial Intelligence*. Berlin, Germany: Springer.

POIBEAU T. (2003). *Extraction automatique d'information : du texte brut au web sémantique*. Paris, France: Hermès.

RILOFF E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, p. 1044–1049.

SHINYAMA Y. & SEKINE S. (2006). Preemptive Information Extraction using Unrestricted Relation Discovery. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, p. 304–311, New York City, USA: Association for Computational Linguistics.