Elaboration de techniques d'analyse adaptées à la construction d'une base de connaissances.

F.ROUSSELOT Université de Strasbourg II ERIC-GRILL d402fr@frccsc21.earn
B.MIGAULT ENSAIS Strasbourg, P.IGOT Université de Strasbourg II

Introduction

Le problème de l'acquisition de connaissances pour constituer une base de connaissances ou modéliser le comportement d'un expert nécessite souvent le traitement de données verbales. Les techniques courantes en linguistique computationnelle ne semblent pas adaptées: il est irréaliste d'écrire un analyseur spécialisé pour chaque sous-domaine considéré (chaque système expert) (cf SAGER et al. 87). Ces données verbales sont souvent incorrectes (retranscriptions de conversations).Les approches statistiques (ZERNIK et al.90) sont inadaptées.

Notre objectif est d'accélérer le déchiffrage du domaine en procurant, sur une station de travail dédiée à la modélisation, les outils nécessaires au traitement de la langue même, afin d'assister le cogniticien dans la première phase d'analyse d'un domaine. Comme d'autres chercheurs (WROBEL 90) (GRUBER 90), nous pensons qu'il faut développer un processus coopératif entre l'utilisateur et le système qui doit prendre en charge le plus possible la vérification de la cohérence des entrées, qui doit aider à poser de bonnes questions à l'expert, qui doit suggérer des choix d'interprétation, proposer des concepts... A la différence de ces auteurs, nous envisageons cette coopération dès l'étape d'analyse des données textuelles, sans prévoir un prétraitement manuel de la langue comme (MÖLLER 88) et (KERSTEN 86), sans nous restreindre à un domaine particulier comme (SCHMIDT et al 91) (JANSEN-WINKEL 88).

Nous basant sur les régularités du langage scientifique dans la situation de recueil d'expertise, nous élaborons des outils généraux, c'est à dire indépendant du domaine et d'un type de problème particulier. Nous étudions surtout des domaines où modéliser des actions est important: planification, conception, conception de plans, conception d'expériences (JELTSCH et al. 89), mais aussi des problèmes de type diagnostic.

Cette recherche s'inscrit dans le contexte de la modélisation de connaissances (KEITH et al. 90), nous n'aborderons ici que l'aspect traitement de la langue.

Ce processus consiste ici aussi en la création d'une certaine représentation du texte. Celle-ci doit être adaptée à son utilisation future: la compréhension d'un domaine et de certains problèmes à résoudre dans ce domaine. Newell, (NEWELL 82) parle de trois abstractions à extraire des données, ce sont: les buts de l'agent intelligent, les actions dont il est capable, et les connaissances qui lui permettent de réaliser ces buts. Les expressions d'actions rencontrées dans les données linguistiques sont donc déterminantes pour l'analyse de l'expertise. La représentation doit avoir une certaine adéquation psychologique (CARD et al. 83)(VALLACHER 87) (KLAUS et al. 72), opératoire (MAHLING 90) cf GOMS (CARD et al. 80) et linguistique

(LOFFLER-LAURIAN 82 et 83). Elle doit être structurée, permettre l'expression d'attributs qui décrivent les objets, par exemple, les liens partie-tout (WINSTON et al. 87) (IRIS et al 88) ou qui décrivent les actions (VOGEL 88).

II Le modèle

Nous ne donnons ici qu'un aperçu de notre modèle nécessaire à la compréhension de notre approche. Nous distinguons quatre concepts de bases que nous définissons rapidement de façon informelle ci-dessous. Il s'agit des objets, des actions, des contraintes et des règles qui permettent essentiellement de savoir quand déclencher les actions. A chaque concept, est attachée une structure.

Les **objets** concernent les entités du domaine qui peuvent intervenir dans les actions des problèmes considérés. Les objets classiquement forment une hiérarchie de classes et ont des instances: les objets spécifiques. Aux objets sont attachés des attributs qui suivant les cas servent à décrire des propriétés de l'objet (relations unaires) ou des relations avec d'autres entités du modèle. Certains attributs structuraux font partie du modèle et sont prédéfinis: "est-un" et "instance_de". Les attributs "intervient_dans ", "affecté_par" participent à la description intensionnelle des objets, ils relient des objets à des actions nous l'expliquons plus loin. Pour une classe donnée (un concept), un certain nombre d'attributs sont obligatoires dans un domaine donné (inaliénables)[1], d'autres non. Certains attributs sont valués, par exemple "longueur". Chacun d'eux a une unité associée et éventuellement un intervalle de définition, un type etc... D'autres attributs sont bien sûr propres au domaine, par exemple pour la classe enzyme "le site de restriction" associé[2]. Ceux-ci sont importants, car liés en général aux actions: ils apparaîtront dans les contraintes d'actions ou dans les règles.

La description des **actions** est basée sur l'hypothèse que le temps est discrétisable et qu'une action peut être décrite par un état initial et un état final. Nous distinguerons plusieurs types d'actions: les **actions simples** considérées par l'expert comme des opérations élémentaires non décomposables, les **actions complexes** qui peuvent être décomposées en une suite (ou plus rarement une structure) d'actions simples.

Une action possède les attributs "agent" (la plupart du temps l'expert) qui décrit celui qui fait et contrôle l'action, "objet_affecté" , "état_final" et éventuellement "objet_ap" (objet présent en arrière plan dont la présence conditionne le déroulement de l'action: instrument par exemple, le catalyseur dans une réaction etc...) et des attributs hiérarchiques "est-un" pour relier une action à une action générique ou "liste_des_sous_actions" pour une action complexe.

Les verbes correspondent souvent à des actions complètement définies (cf ci-dessus), ils correspondent parfois à des actions incomplétement spécifiées, communément appelées plans. Cette notion n'est pas rigoureusement définie dans la littérature, nous précisons donc notre terminologie ici.

Une **action-objectif** fait référence soit à un objectif, soit à un enchaînement de tâches abstraites. Ainsi, une action-objectif peut être définie uniquement par l' énoncé de son objectif ou par une liste de méthodes pour l'atteindre. Chaque méthode est appelée **plan**: il s'agit d'une description des actions à entreprendre pour espérer arriver à réaliser l'objectif. Cette description est effectuée en terme d'autres actions-objectif: le problème de l'expert se ramène souvent à construire un plan complétement

[1]Nécessaire à la description d'un objet dans le cadre du problème envisagé dans le domaine.

[2]Un enzyme est un outil destiné à couper une molécule d'ADN, le site de restriction est la description de la séquence d'ADN qu'il coupe (de longueur 5 ou 10 bases).exemple CGTCA

spécifié ou à adapter un plan existant (FRIEDLAND 85). Il est souvent difficile de déterminer si un verbe dans un texte peut représenter une action ou une action-objectif, il faut souvent attendre pour pouvoir décider de quelle type d'entité il s'agit.

Une action-objectif est définie par l'attribut état_final qui décrit l'objectif (qui est en somme un but à atteindre) et par l'attribut "liste_des_plans", il possède les attributs "agent", "état_initial" (éventuellement vide) et "objet_ap" (idem).

On remarque que les plans sont comparables à des actions complexes, mis à part le fait qu'ils sont incomplétement spécifiés: ils ont donc les mêmes attributs que les actions complexes: l'attribut liste_des_sous_actions renvoie sur une structure bâtie sur des actions-objectif. Un plan est comparable à un programme informatique et comporte éventuellement boucle et "instructions" de contrôles.

Nous avons encore deux types d'entités que nous ne détaillons pas les **règles** qui relient les méthodes aux actions-objectif et les **contraintes** liées aux actions et à la description des objets.

III Analyser

Sans insister sur la validité du modèle proposé, posons le problème du traitement linguistique.

1) la langue scientifique

Sur le plan lexical, on distingue dans le discours scientifique trois niveaux d'emploi des lexèmes

- les lexèmes employés dans le même sens que celui qu'ils ont dans le langage courant ;
- les lexèmes spécifiques au domaine
- les lexèmes appartenant au lexique courant, mais détournés de leur sens.

Remarquons qu'un même terme peut être employé dans le discours scientifique à la fois dans son sens courant et dans un ou plusieurs sens spécifiques (par exemple le cas du verbe *transformer* dans le domaine du génie génétique a un sens général et deux sens spécifiques). Le système que nous développons doit aider l'utilisateur dans la construction d'un lexique du domaine en proposant des entrées multiples et en repérant les différents emplois dans les textes. Ceux-ci dépendent des types d'arguments des diverses occurences qui doivent être analysés et structurés avec soin.

Dans la situation envisagée, nous voulons profiter de l'emploi fréquent par l'expert des constructions toutes faites, de sa façon standard d'organiser et de présenter ses propos qui relève d'une rhétorique propre à la langue et à la pratique scientifique.

2)méthode

Nous avons constaté deux choses - les moyens de traitements "faibles" de la langue (NORVIG 89) sont plus adaptés à la généralité et aux extensions, que les systèmes forts, du type analyseurs exhaustifs de la langue - il n'y a pas de frontière entre les niveaux morphologie, syntaxe et sémantique, notre compétence linguistique est basée sur des schémas constructifs mêlant ces niveaux (LANGACKER 87). Aussi, avons-nous choisi de rechercher des schémas morpho-syntaxiques caractéristiques qui sont en principe réutilisables dans tout domaine scientifique. Ils servent à écrire des règles destinées à associer des mots ou des groupes de mots aux structures de la représentation choisie. L'interprétation de ces règles ne nécessite qu'un analyseur trés simple.

Nous avons d'abord relevé un certain nombre de critères pertinents comme par exemple: la "détermination", l'usage de l'infinitif, l'emploi de certains verbes spécifiques pour l'expression des règles (voir les exemples plus loin). Des hypothèses de règles ont été ensuite élaborées à partir de différents corpus scientifiques représentatifs (IGOT 91), puis vérifiées sur l'intégralité de ceux-ci. Elles devront bien entendu être validées sur des corpus plus importants, mais on peut déjà avoir une certaine confiance en elles.

Nous présentons ici un certain nombre de règles de façon informelle pour montrer les possibilités de notre approche. Les schémas sont éventuellement accompagnés de contraintes.

Quelques exemples:

objets

attribution d'une propriété

ex1 "le/la <nom> de <objet> " => <nom>=propriété

contraintes liées à ex 1: objet est un objet du domaine, nom n'est pas une propriété répertoriée, n'est pas dans la liste (groupe, ensemble, extrémité...), nom n'est pas un nom verbal

exemple: "la couleur de la solution"

constitution d'une classe

ex2 "il y a, il existe" n < objet O>-s ["différents"]: <nom1>, <nom2>,..<nom$_n$>

la classe objet O est constituée de ...

exemple: "il y a quatre bases: ...."

détermination d'une sous-classe

ex3 <nompropre> "est un" <objet O>[<attributs>] => <nom> =<sous-classe de objet O>

ex4 "un" <nom> "est un" <objet O> [<attributs>] => <nom> =<sous-classe de objet O>

ex5 "les" <nom>-s "sont des" <objet O>-s [<attributs>] => <nom> =<sous-classe de objet O>

(l'écriture <obj O> fait référence à une classe d'objets O répertoriée)

exemple "les vecteurs sont des molécules d'ADN"

relation partie-tout

elle est surtout basée sur les occurences de "avoir", de "posséder" exemple

ex 6 <objet> "a un/des"<nom>-s [+Sadj] => <nom> désigne une partie de <objet> déjà répertorié comme objet du domaine.

En général, on ne se contente pas de définir une partie, dans le même énoncé on la qualifie d'où la possibilité du syntagme adjectival Sadj.

Exemple "les molécules ont des extrémités franches"

D'autres verbes sont plus productifs: "faire partie de", "être composé de" etc... ils indiquent une relation *partie-tout* qui permet de conclure que les deux arguments sont des objets.

liaisons entre les objets et les actions

Il est important de savoir si un objet peut intervenir dans un action ou être affecté par une action.

ex7 "on peut" <infinitif><objet>

<objet> "peut" <infinitif passif>

<objet> "sert à" <infinitif>

on détecte ici la présence d'actions (les infinitifs).

exemple: on peut couper la molécule..."

attributs quantifiés

Les schémas sont nombreux, ils sont en général basés sur des prépositions : "de...à" etc... des nombres, sur des noms de propriétés mesurables: " d'une longueur de X mètres ", ainsi que sur des unités.

synonymie

ex8 "On appelle <nom> ,un <nom>" le dernier nom étant non suivi d'un syntagme adjectival .

actions

Elles doivent être tout d'abord repérées: un prétraitement morphologique basé sur certains suffixes connus peut fournir bon nombre de candidats à être une action (mais pas tous, exemple: electrophorèse).

règles basées sur des verbes opérateurs (LOFFLER-LAURIAN 82 et 83)

ex9 "faire/réaliser/effectuer un/le "<nom> ["de"<objet>] confirme un candidat .

autre utilisation possible: quand on pense être sûr d'avoir répertorié tous les objets du domaine: si <nom> n'en fait pas partie, c'est certainement une action. exemple: "faire une electrophorèse".

III programmation

1) schémas.

Nous allons dans la suite en donner quelques exemples démonstratifs et en profiterons pour

introduire le langage d'écriture de ces schémas qui utilisent un nombre minimum de critères identifiés.

l' exemple ex1 s'écrit

< Art: def >< Mot: nom: non domobj: non nomverb: non listexp1>de<Gnom: domobj >

=> <Mot> = < propriété>

Un schéma est destiné à faire une sorte de pattern-matching, il comporte des chaînes de caractères, des suffixes et des catégories spécifiant des ensembles de mots possibles.

Les doubles points servent à écrire les contraintes qui sont morphologiques, syntaxiques, qui portent sur la détermination, ou qui relèvent de l'état d'élaboration de la base: par exemple, dans l'exemple, domobj signifie que le Gnom doit référer à une entité déjà répertoriée comme un objet du domaine.

Gnom est une information obtenue de l'analyseur, il signifie que le programme doit analyser ici un groupe de mots comme groupe nominal.

On peut interpréter ces symboles comme des prédicats à vérifier: on dispose d'une liste de prédicats prédéfinis

- portants sur les mots et relevant de l'analyse: Mot, Art (teste un article), Nompropre (teste un nompropre), Inf(pour infinitif), Gn (teste un groupe nominal) etc...
- morphologiques: suffaction (suffixes des actions: ion, age, ment)
- portants sur des traits syntaxiques: def (pour défini), sing, plur etc.
- portants sur l'état d'avancement de la base: candaction (candidat à être une action), objdom (objet répertorié comme objet du domaine), actiondom (action du domaine) etc...

On peut utiliser "non" pour signifier qu'un prédicat ne doit pas être vrai, le double point dans un élément de schéma entre "<" et ">" joue le rôle du connecteur logique "et".

Dans le cas de succès à tous ces tests, la règle s'applique et l'on déduit que le premier mot est un nom de propriété.

L'exemple 1 caractérise un énoncé de description d'objet. voir paragraphe précédent.

exemple2 : "la <Mot: suitcar-"ion" > de <Gnom : domobj > "

le "matching" aura réussi si la phrase contient une séquence de mots du type

"la xxxx-ion de l'enzyme"

<mot: suitcar-"ion" >spécifie n'importe quel mot avec la contrainte qu'il soit constitué par une suite de caractères (suitcar est un mot-clé du langage) suivie du suffixe ion. (le signe "-" signifie "suivi de")

exemple3: < Mot: domobj: non candaction > signifie que le mot attendu à cette place du schéma est un mot qui désigne un objet du domaine mais pas un candidat à être une action.

On a l'a vu, toutes les règles ne permettent pas forcément de conclure immédiatement à une décision unique quant à un syntagme particulier, cependant elles peuvent servir à restreindre le champ des recherches. Un groupe de mots peut à un endoit être repéré comme action ou classe et à un autre endroit par une autre règle comme classe ou objet, le système que nous implémentons en tient compte.

2) implémentation

L'implémentation est actuellement en cours, avec pour objectif de tester ces règles, de les classer selon leur vraisemblance et de les intégrer au système existant. Le but est de diminuer le plus possible l'intervention humaine dans la première phase de l'acquisition, l'adjonction future d'un gestionnaire d'hypothèses est prévue.

Le programme dispose d'informations préalables. Un dictionnaire des mots non spécialisés est donné, il contient les mots de la métalangue qui servent à définir des concepts exemple : "se composer de", "être formé de"... des notions générales qui peuvent être utiles dans le domaine scientifique: taille longueur, extrémité...Le système dispose d'un analyseur morphologique qui traite: la conjugaison des verbes, les pluriels des noms, des adjectifs.

L'utilisateur du langage peut se définir des prédicats qui généralisent des constructions. Par exemple dans le schéma suivant

< Gnom:domobj> avec un (e) <Mot: propriété> <delta >

delta recouvre toute une série de schémas associés à des quantités : "de <Nombre> à <Nombre>", "entre <Nombre> et <Nombre>" etc...

exemple " le cuivre utilisé avec une pureté de 99%."..

Métaopérateurs

Ceux-ci permettent de gérer les contraintes qui lient les entités des schémas: un schéma peut avoir une interprétation totalement déclarative, ç.à.d. si dans une suite de mots pour un schéma donné, **un seul** prédicat est non calculable, il est déduit.

Exemple: du schéma précédent, si on sait que le groupe nominal est un objet du domaine alors on peut en déduire que le mot qui suit "avec un(e)" est une propriété, de même si on sait que le second mot est une propriété, on peut en déduire que le premier est un objet du domaine.

écriture: decl (schéma)

Si le schéma n'est pas déclaratif, on signale par un ? suivant la propriété à déduire, la conclusion potentielle du matching du schéma

Exemple

"on peut effectuer le" <Mot: suitcar-"age": candidaction: domaction?> de <Gnom: objdom>" suitcar teste une suite decaractères quelconque.

Lorsque le matching réussit ici on en déduit que le mot est la dénomination d'une action du domaine.

Conclusion

Les premiers résultats sont encourageants, l'approche semble prometteuse. Nous avons rassemblé plus de cent schémas que nous expérimentons. Une étude est actuellement en cours sur une intégration ergonomique de ces schémas sur la station de travail d'acquisition de connaissances. Le fait que les structures possèdent des références croisées, rend possible certaines vérifications de cohérence, la détermination de questions pertinentes à poser à l'expert, la détections d'anomalies: notions inutiles etc...

Nous développerons par la suite deux axes: l'apprentissage des schémas (apprentissage par l'exemple) et la recherche, dans les données verbales, d'indices de la présence d'opérateurs d'inférence ou de tâches génériques (CHANDRASEKARAN 86) en généralisant l'utilisation de champs sémantiques liés à ces tâches (MÖLLER 88 déjà cité). On peut déduire, par exemple de la présence du mot "norme" la présence d'un ensemble de "paramètres" à "comparer" à celle-ci, ou à partir d'occurrences de mots tels "trier" "caractériser" "identifier" détecter la présence de primitives d'inférences[1]. Notre but est d'aboutir à une méthode directive (dirigée par les données verbales) d'analyse de domaine; en effet, la plupart des méthodes actuelles souffrent d'un manque de directivité.

Nous pensons que les résultats actuels ont déjà une portée générale et seront réutilisables dans des contextes dépassant le cadre du recueil d'expertise. Par exemple, il est envisageable de réutiliser la bibliothèque de schémas pour produire des textes (explicatifs) à partir d'une base de connaissances. Il est, en outre, vraisemblable que cette approche, dans son principe, est indépendante d'une langue particulière.

Bibliographie

BOESH "Kultur und Handlung" Ed. Huber Bern 1980

CARD S.K., MORAN T.P,. NEWELL A. "Computer text-editing: an information-pro-

[1] Grosso modo ce terme désigne toute manipulation conceptuelle entrant dans un processus de résolution de problème.

cessing analysis of a routine cognitive skill" *Cognitive psychology N°12* p 32-74. 1980

CARD S.K., MORAN T.P,. NEWELL A. "The psychology of Human-Computer interaction" Hillsdale Lawrence Erlbaum 1983

CHANDRASEKARAN B. "Generic tasks in knowledge-based reasoning: High-level building blocks for expert system design" *IEEE Expert* 1(3) p 23-29 1986

FRIEDLAND P., IWASAKI Y. "The concept and implementation of skeletal plans "*Journal of Automated Reasoning 1* p.161-208 1985

IGOT P. "Eléments d'analyse linguistique en vue d'une modélisation des connaissances sur ordinateur" mémoire de DEA 110p public.interne GRILL 1991

IRIS M A, LITOWITZ B.E, EVENS M "Problems of the part-whole relation" in Relational models of the lexicon Ed M. W EVENS p262-288 Cambridge University Press 1988

JANSEN-WINKELN R.M. "WASTL: An approach to Knowledge Acquisition in the Natural Language Domain" Bericht N° 48 Universität des Saarlandes Saarbrücken 1988

JELTSCH J.M., KEITH B., MIGAULT B., ROUSSELOT F. "Reasonning about manipulation of DNA molecules" Simposium International de Intelligencia Artificial Mexico 1989

KEITH B., MIGAULT B., ROUSSELOT F. " A Methodology of Knowledge Acquisition" Proceedings of AIMSA 90 Belgrade 1990

KERSTEN M.L. "A Conceptual Modeling Expert System" Report CS R8518 Centre for Mathematics and Computer Science Amsterdam 1986

KLAUS G., BUHR M. "Philosophisches Wörterbuch" VEB Deutscher Verlag der Wissenschaften Berlin ex GDR 1972

LANGACKER R. W. "An introduction to cognitive grammar" *Cognitive Science 10* p1-40 1986

LOFFLER-LAURIAN A.M. "Être dans quelques textes en langue naturelle" *Revue de Linguistique Romane N°181-182* Strasbourg p121-157. 1982

LOFFLER-LAURIAN A.M. "Faire et ses quasisynonymes dans les discours scientifiques *Etudes de Linguistique Appliquée (nlle série) N°51: "Les discours scientifiques"* Paris. Ed Didier-Erudition p93-103 1983

MAHLING D.E. GROFT W.B. "Relating Human Knowledge of Tasks" in *Foundations of knowledge Acquisition.* Knowledge Based System Vol 4 Ed Boose J.H Gaines B.R. Academic Press 1990

MÖLLER Jens-Uwe "Knowledge Acquisition from texts" EKAW 88 ST Augustin p25-1, 25-16 1988

NEWELL " The knowledge level " *A.I. N°18* p87-127 1982

NORVIG P. "Marker passing as a weak method for text inferencing" *Cognitive Science 13* p.569-620 1989

SAGER N., FRIEDMAN C., MARGARET S., LYMAN MD. "Medical Language Processing of Narrative Data". Ed Addison Wesley 1987

SCHMIDT G., SCHMALHOFER F "Situated text analysis with COKAM+" to appear in EKAW 91

VOGEL C. "Génie Cognitif" Ed Masson .1988

WINSTON M, CHAFFIN R, HERRMAN D "A taxonomy of part-whole relations" *Cognitive Science 11* p.417-444 1987

WROBEL S. "Design goals for a sloppy modelling system" in *Foundations of Knowledge Acquisition.* Knowledge Based System Vol 4. Ed Boose J.H. Gaines B.R. Academic Press 1990

ZERNIK U., JACOBS P. "Tagging for Learning: Collecting Thematic Relations from Corpus" Proceedings of COLING 90 . 1990