Le raisonnement à partir de cas pour l'identification de rôles sémantiques dans des énoncés en langue naturelle

Fairouz Chakkour, Amedeo Napoli, Yannick Toussaint

Le Raisonnement à partir de cas pour l'identification de rôles sémantiques dans des énoncés en langue naturelle

Fairouz CHAKKOUR, Amedeo NAPOLI, Yannick TOUSSAINT

Équipe ORPAILLEUR,
LORIA – UMR 7503,
B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France
Fairouz.Chakkour@loria.fr, Yannick.Toussaint@loria.fr, Amedeo.Napoli@loria.fr

Résumé

Les énoncés en langue naturelle dans les domaines techniques présentent des constructions syntaxiques récurrentes. Nous proposons de mettre en œuvre un système de raisonnement à partir de cas pour nous permettre de passer de l'analyse syntaxique d'une phrase à sa représentation conceptuelle.

Mots clef : raisonnement à partir de cas, traitement automatique de la langue, fouille de données textuelles, analyse conceptuelle de phrases, extraction d'information, analyse sémantique de phrases.

1 Objectifs et Motivations

Le but de notre recherche est de concevoir un système capable de produire une interprétation d'un énoncé en langue naturelle qui puisse être utilisée dans un processus de fouille de données textuelles. Nous nous intéressons à l'exploitation et à l'application des principes du raisonnement à partir de cas pour aboutir à cette interprétation. L'analyse d'un énoncé repose sur plusieurs types de connaissances : un thésaurus, une base de cas et un ensemble de schémas de phrases. Les données en entrée sont des phrases associées à un premier niveau d'analyse syntaxique.

Notre motivation repose sur le fait que, dans le cadre du traitement de textes scientifiques et techniques, il est possible d'observer un certain nombre de régularités dans les phénomènes de langue. De plus, l'interprétation d'un énoncé doit reposer sur des connaissances du domaine. Un système à base de cas doit nous permettre la prise en compte de ces spécificités. Nous commençons ce papier par la présentation de l'idée générale de l'approche illustrée par un exemple introductif qui explicite l'entrée et la sortie du système (cf.section2). Nous décrivons ensuite le système exploitant le raisonnement à partir de cas (cf.section 3) par son architecture globale et par les notions générales du RàPC, avant de décrire les connaissances utilisées par ce système : la base de cas, le thésaurus et le schéma de phrases. Dans la dernière section (cf.section 4) le processus de REMÉMORATION sera présenté en donnant un exemple. Enfin, une conclusion où est présenté un certain nombre de perspectives de recherches pratiques et théoriques termine le papier.

2 De l'analyse syntaxique à l'analyse conceptuelle d'un énoncé

L'interprétation d'un énoncé consiste à identifier, pour chacun des actants, agent et objet, le rôle qu'il joue dans la phrase et à identifier également la relation que le verbe établit entre ces actants, ou l'événement communiqué par le verbe. En conséquence, cela revient à produire un niveau d'abstraction supplémentaire par rapport à la syntaxe. Ainsi, dans un énoncé comme *Marie possède un ordinateur*, savoir que la relation entre *Marie* et l'*ordinateur* est une relation de POSSESSION est plus informatif que la simple décompositon syntaxique selon laquelle *Marie* est le sujet du verbe *posséder* et l'*ordinateur* est le complément d'objet du même verbe. De même, dans un énoncé comme *Jean possède un cerveau* , la relation entre *Jean* et *cerveau* est une relation de composition, appelée ici `partie_de`, exprimant le fait que le *cerveau* est une partie de *Jean*. Cela malgré le fait que *Jean* et le *cerveau* ont les mêmes rôles syntaxiques par rapport au verbe *posséder* que *Marie* et l'*ordinateur*, respectivement ; sujet et complément d'objet.

Nous supposons disposer comme données initiales d'un énoncé étiqueté morpho-syntaxiquement et découpé en syntagmes nominaux et verbaux, comme le proposent [2] et [9]. Ainsi, l'énoncé : *Jean possède un cerveau* est associé à la structure syntaxique suivante, qui représente l'entrée du système :

```
Jean : GN / suj:v posséder
possède : GV
un cerveau : GN / c.o.d.:v posséder
```

L'interprétation de cette phrase, qui est la sortie du système, sera représentée par une structure qu'on appelle concept, comme suit:

```
l1. c - posséder - partie_de
l2. relation: partie_de
l3. agent: lex: Jean
l4. agent: rsyntax : sujet
l5. agent : rôle : objcomposé
l6. agent: type : objanimé
```

```
l7. objet: lex : un cerveau
l8. objet: rsyntax : c.o.d.
l9. objet: rôle : composant
l10. objet: type : organe
```

`c - posséder - partie_de` exprime le fait que la phrase contient le verbe *posséder* comme verbe principal, (ligne l1.), qui communique une relation de composition, nommée `partie_de`, entre l'agent et l'objet (ligne l2.). Dans ce concept, "Jean" forme la facette lexicale, `lex` de l'agent (ligne l3.). Son rôle syntaxique, `rsyntax`, est le sujet du verbe. Sémantiquement l'agent est un objet composé, `objcomposé`, et sa classe dans le thésaurus est `objanimé`. De la même façon l'objet est représenté par ces quatre facettes: un cerveau, c.o.d., composant, organes.

Cette structure est inspirée des travaux en extraction d'information à partir de textes en langue naturelle, notamment les travaux de G.Lapalme et d'E.Riloff, [4] et [8]. Ils utilisent des structures semblables dans le processus d'extraction de l'information.

Pour notre compte, nous nous intéressons aux énoncés constitués par des phrases simples. Néanmoins, nous envisageons, à court terme, l'extension de nos travaux à des énoncés au sens plus large du terme.

3 Le système exploitant le raisonnement à partir de cas

Dans notre approche, l'analyse de phrases est dependente d'exemples déjà analysés existants dans une base de connaissances appelée base de cas (cf. section 3.3). Cette approche est nommée **analyse de phrases à partir de cas**. Celle-ci utilise des connaissances extraites de la base de cas, du schéma des phrases et du thésaurus pour construire l'analyse. Cette approche repose sur l'idée suivante : si deux phrases sont "semblables", leur analyse est également "semblable". Ainsi, si l'analyse d'une première phrase est connue, l'analyse d'une phrase semblable peut être obtenue en réalisant une adaptation de la première. De nombreux travaux dans le raisonnement à partir de cas sont consacrés au traitement automatique du langage. Ils s'intéressent à l'exploitation et à l'application des principes du RàPC à la traduction automatique [10] [7], ainsi qu'à l'analyse syntaxique des phrases [5], à l'analyse de textes [1], à la classification de textes [8] ou encore à la recherche d'information [6].

Dans le but de produire la représentation conceptuelle d'une phrase, on adopte le modèle du raisonnement à partir de cas fondé sur des principes de classification hiérarchique et des chemins de similarité. Ce modèle est présenté dans la thèse de J.Lieber [3].

3.1 Architecture globale du système

L'architecture globale du système, illustrée dans la figure suivante (cf. fig.1), montre le déroulement du processus en prenant en entrée une phrase déjà étiquetée et découpée en syntagmes et en produisant comme sortie du système le concept correspondant à cette phrase.

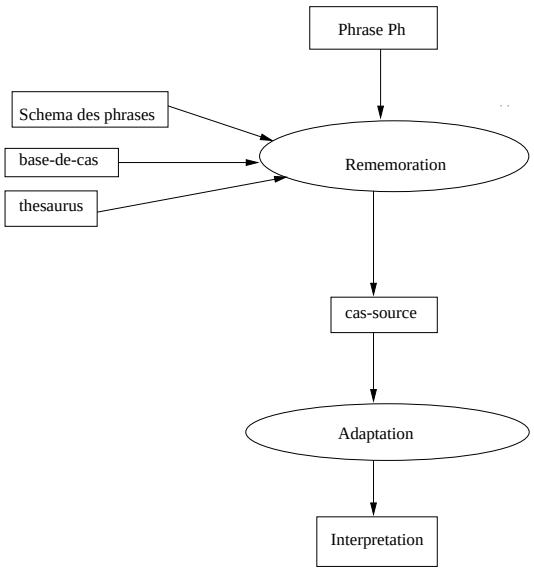

FIG. 1 – *Architecture globale du système*

Nous allons expliquer les étapes de fonctionnement du système dans ce qui suit. Nous faisons un rappel des notions générales du raisonnement à partir de cas, avant de décrire brièvement les connaissances utilisées par notre système : la base de cas, le thésaurus et le schéma de phrases. Ces connaissances seront exploitées dans l'étape de REMÉMORATION (cf.section 4). En revanche, nous n'aborderons pas dans ce papier les notions d'adaptation et de chemins de similarité.

3.2 Notions générales sur le RàPC

La base de cas est la base de connaissance centrale d'un tel système. Un cas `C` est considéré classiquement comme la donnée d'un problème `P` et d'une solution `Sol(P)` à ce problème. Un cas de la base de cas est appelé cas source, `cas-srce` , tel que `srce` est l'énoncé du problème source. Étant donné un problème à résoudre appelé problème cible, un système à base de cas a pour tâche d'exploiter la base de cas pour résoudre le problème cible. Ce raisonnement s'effectue en deux phases: la REMÉMORATION et l' ADAPTATION. La remémoration a pour objectif de chercher dans la base de cas un cas source `cas-srce` tel que `srce` est similaire au problème cible. L'adaptation a pour tâche de construire une solution `Sol(cible)` au problème cible en s'appuyant sur la solution du problème source.

3.3 La base de cas

Dans notre approche, le problème source est une phrase source, `phsrce`, et un problème cible est une phrase cible, `phcible`. Dans la base de cas, le cas associé à une phrase source donnée est un couple contenant la **représentation** de la phrase `phsrce`, et l' **interprétation** de la phrase `phsrce`, `cas-srce = (rep(phsrce), Int(phsrce))`.

La **représentation** de la phrase `phsrce` contient la chaîne des mots de la phrase, l'étiquetage de ces mots, et aussi le découpage de cette phrase en syntagmes. L'**interprétation** de la phrase `phsrce` est le concept correspondant à cette phrase.

Pour mieux définir et comprendre la base de cas ainsi que la suite du processus, on prend l'exemple d'un cas dans la base de cas.

Exemple 3.3:

Phrase source `phsrce`: Marie possède une poupée

Représentation: rep(`phsrce`)

```
Marie: SN/ suj:v.posséder
possède : SV
un jouet : SN/c.o.d.:v. posséder
```

Intérprétation: Int(`phsrce`)

```
c - posséder - possesion
relation: possession
agent: lex: Marie
agent: rsyntax : sujet
agent : rôle : possesseur
agent: type : objanimé
objet: lex : une poupée
objet: rsyntax : c.o.d.
objet: rôle : objpossédé
objet: type : objphys
```

Avant d'aborder des situations plus complexes, nous avons choisi de commencer par l'analyse de phrases simples. En conséquence, les phrases choisies pour être représentées dans la base de cas ont une structure syntaxique élémentaire, et les groupes nominaux ne contiennent qu'un déterminant et qu'un seul nom, et les groupes verbaux qu'un seul verbe. Ces phrases sont spécifiques, dans le sens où on considère comme phrase source *Marie possède un ordinateur* et non pas une phrase plus générique comme pourrait l'être *X possède Y*. Ce sont des phrases "artificielles" constituées dans le but de rendre plus facilement compréhensible notre démarche.

3.4 Thésaurus

Le thésaurus représenté par la figure (cf. fig.2) est une hiérarchie de termes. Nous l'utiliserons comme source de connaissances en considérant que le thésaurus définit une structure dans laquelle un terme est associé à une classe plus spécifique que ses pères et plus générique que ses fils.

Ce thésaurus nous sert à trouver une certaine similitude entre les noms. Les noms qui sont plus proches dans la hiérarchie sont similaires. Autrement dit, deux noms séparés par une distance plus courte, sont plus similaires que deux noms séparés par une distance plus longue.

Étant donnés deux noms N_1 et N_2, la distance entre eux, notée $dis(N_1, N_2)$, est le coût du chemin entre ces deux noms. Ce coût est l'addition des coûts des arcs parcourus dans le thésaurus pour aller de N_1 à N_2. Autrement dit le coût des transformations de généralisation et des spécialisation efféctuées.

Ce coût ne peut être fixé indépendamment du verbe et du domaine en question.Cependant nous le fixons à 1 pour tous les arcs du fait que la déterminaison d'un coût "réel" importe peu ici.

À titre d'exemple, étant donné les deux noms dans le thésaurus, **Marie** et **Claire**, la distance
dis(**Marie**, **Claire**) = coût arc(**Marie**, femme) + coût arc(femme,**Claire**) = 1 + 1 = 2
Les coûts des arcs ainsi que l'heuristique du calcul de la distance sont des points que nous allons explorer dans un avenir proche.

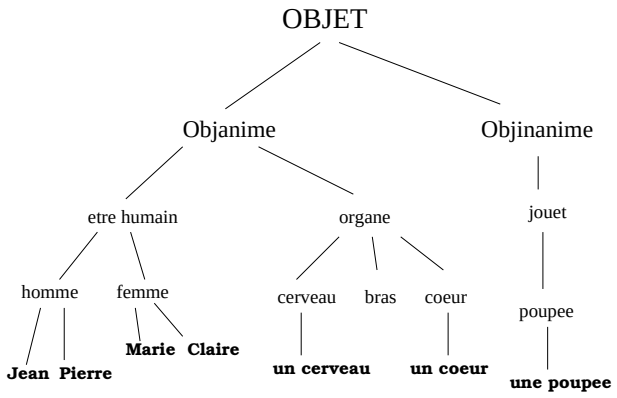

FIG. 2 – *Le thésaurus*

3.5 Le schéma des phrases

Comme nous l'avons mentionné précédemment, les cas sources sont des phrases réelles, et elles correspondent à un faible niveau de généralité. En conséquence, pour permettre d'accéder à un cas source dans la base-de-cas, on a besoin de les indexer. Nous supposons qu'à chaque cas source `cas-srce = (rep(phsrce), Int(phsrce))` est associé un index `ind(phsrce)` qui est une phrase plus générale que la phrase `srce`. Le schéma des phrases est une hiérarchie des index, (cf. fig.3), la racine dans cette

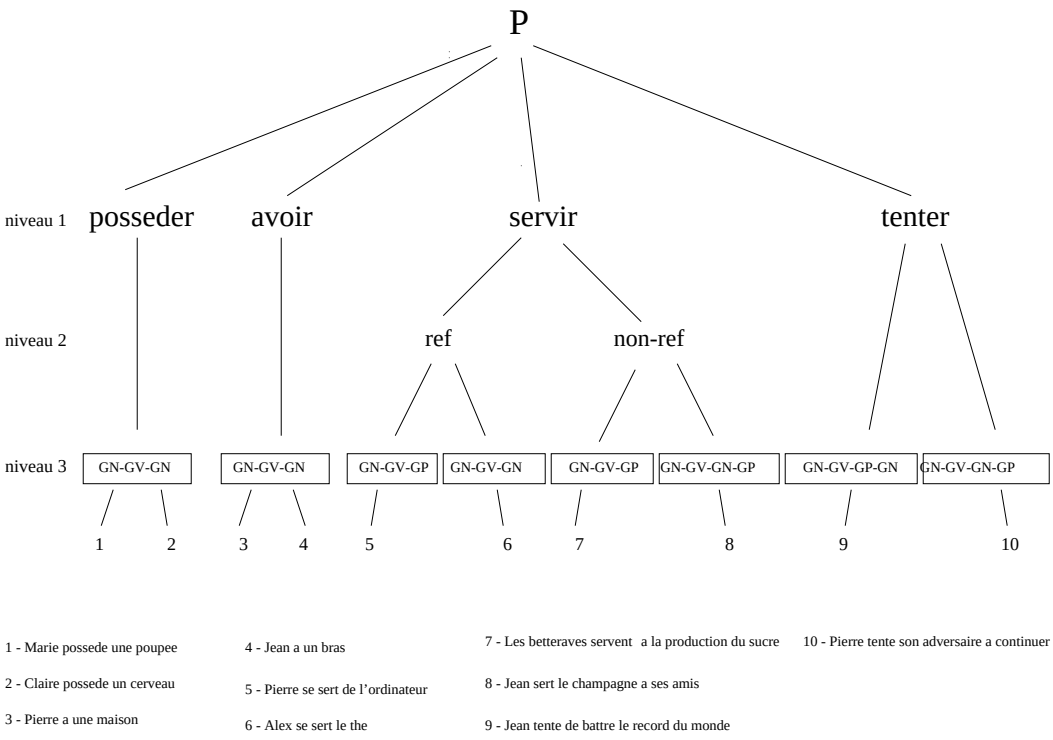

FIG. 3 – *Le schéma des phrases*

hiérarchie est la phrase générique, les classes du premier niveau de la hiérarchie sont construites en fonction du verbe de la phrase. Par exemple dans la figure (fig.3), on a quatre classes de phrases au premier niveau en fonction de quatre verbes : posséder, avoir, servir, tenter. Chaque classe regroupe les phrases qui contiennent le verbe correspondant. Ainsi les classes du troisième niveau de la hiérarchie sont construites en fonction de la structure syntaxique de la phrase. Par exemple dans la figure (fig.3) l'index GN-GV-GN, qui représente une sous classe de la classe "posséder", regroupe les phrases qui contiennent le verbe *posséder* et qui ont un schéma syntaxique de la forme "groupe nominal - groupe verbal - groupe nominal". Pour certaines classes de phrases, on envisage deux sous-classes, "ref" et "non-ref", dans un deuxième niveau de la hiérarchie. Cela dépend si le verbe en question est de type réflexif, comme le verbe "se servir" dans la phrase 5 (cf. fig.3), ou non-réflexif, comme le verbe "servir" dans la phrase 8 (cf. fig.3).

L'index peut pointer sur plusieurs cas à la fois. Par exemple, les deux phrases "Marie possède un jouet" et "Claire possède un cerveau" ont le même index. Cela est illustré dans le schéma des phrases (cf. fig.3).

Étant donnée une phrase cible en entrée, `phcible`, pour chercher la phrase source, `phsource`, qui lui est similaire, on cherche l'index de `phsrce` dans le schéma de phrases. Cela sera effectué par la CLASSIFICATION.

La CLASSIFICATION est le processus qui permet la recherche de l'index le plus proche de la phrase cible, `phcible`, dans le schéma des phrases [3]. Les informations exploitées dans la CLASSIFICATION sont de nature purement syntaxique.

4 La remémoration

La remémoration est une fonction qui prend en entrée une phrase cible, `phcible`, et retourne un `cas-srce = ( rep(phsrce), Int(phsrce))` qui est un cas de la base de cas. La remémoration doit effectuer quatre types de tâches exécutées de façon séquentielle: une tâche de NORMALISATION, une tâche de CLASSIFICATION, une tâche de CALCUL DE DISTANCE et une tâche de CHOIX.

4.1 Algorithme général de remémoration

L'algorithme général de remémoration consiste en les étapes suivantes:

1. Normaliser la phrase cible `phcible`. La normalisation produit une phrase à la forme active, et dans certains cas une phrase avec un verbe non-réflexif.
2. Classifier la phrase cible, `phcible`, dans la hiérarchie des phrases pour trouver l'index le plus proche `ind-cible`.

3. Si `ind-cible` est l'index d'un seul cas source alors retourner `cas-srce` tel que :
   `cas-srce = cas(index)`
   sinon calculer la distance entre la phrase cible, `phcible`, et chacune des phrases correspondante aux cas source indexés par `ind-cible`.
4. Choisir le cas source le plus proche, qui est celui qui minimise la distance dis_{GN}(`phcible`, `cas-srce`).
   On suppose dis_{GN}(`phcible`, `cas-srce`) = dis_{GN}(`phcible`, `phsrce`)

Nous donnons dans la suite deux définitions précisant les notions de NORMALISATION et de DISTANCE dis_{GN}.

Définition 1 (NORMALISATION)
La tˆache de NORMALISATION *consiste `a effectuer les ´etapes suivantes:*

1. *Identifi er le verbe de la phrase cible* `phcible`.
2. *Si le verbe est `a la forme passive, et si la transformation passif-actif est permise dans la classe du verbe, alors appliquer cette transformation.*
3. *Si le verbe est r´eflexif, et si la transformation r´eflexif-nonr´eflexif est permise dans la classe du verbe, alors appliquer cette transformation.*

Avant de donner une définition de la DISTANCE dis_{GN} entre deux phrases, nous allons rappeler le fait que nous avons choisi de commencer avec de phrases simples (cf. section 3.3). Ainsi, nous supposons que les groupes nominaux à traiter sont de la forme `[Det Nom]` ou `[Nom_Propre]`.
Chaque groupe nominal joue un rôle syntaxique dans la phrase comme étant sujet ou objet du verbe. Ainsi, la notation, $GN_{phsrce}(sujet)$, dénote le groupe nominal de la phrase `phsrce` qui joue le rôle syntaxique du sujet du verbe.

Définition 2 (DISTANCE dis_{GN})
Étant donn´ees deux phrases, `phcible` *et* `phsource`*, qui ont la mˆeme construction syntaxique et le mˆeme verbe, et ´etant donn´e $GN_{phcible}(r)$, un groupe nominal qui appartient `a la phrase cible et ayant le rˆole syntaxique $r$, et $GN_{phsrce}$ un groupe nominal qui appartient `a la phrase source et ayant le mˆeme rˆole syntaxique r, la distance, dis_{GN}, entre* `phcible` *et* `phsource` *est calcul´ee comme suit:*
dis_{GN} (`phcible`, `phsrce`) = Σ_r $dis(GN_{phcible}(r), GN_{phsrce}(r))$
pour l'ensemble des rˆoles syntaxiques instanci´es dans les phrases.
Actuellement, la distance calcul´ee ne prend pas en compte le d´eterminant du groupe nominal. En effet, la distance calcul´ee entre deux groupes nominaux est la distance entre les deux noms ou noms propres des groupes nominaux. Ainsi, le groupe nominal, $GN_{phcible}(r)$, est repr´esent´e par le nom qu'il contient, $N_{phcible}(r)$, et la distance entre deux phrases :
dis_{GN} (`phcible`, `phsrce`) = Σ_r $dis(N_{phcible}(r), N_{phsrce}(r))$
La notion de la distance entre deux noms est pr´esent´ee dans la section (cf. section 3.5).

Exemple 4:
On suppose avoir la phrase cible, `phcible`, *Marie possède un coeur*, et on va dérouler l'algorithme de remémoration.
Supposons que la base de cas contient les phrases suivantes :

```
Marie possède une poupée.
Claire possède un cerveau.

Pierre a une maison.
Jean a un bras.

Jean sert le champagne à ses amis.
Pierre se sert de l'ordinateur.
```

qui sont indexées et organisées dans la hiérarchie des phrases (cf.fig.3).

Le processus de la remémoration se déroule comme suit :

1. Normaliser la phrase:
   (a) Le verbe est *posséder*
   (b) Le verbe n'est pas à la forme passive
   (c) Le verbe n'est pas réflexif
2. Le résultat de la classification de `phcible` dans le schéma des phrases nous donne l'index [GN GV GN] (cf.fig.3)
3. Cet index est l'index des deux phrases:
   `phsrce1`: `Marie possède une poupée.`
   `phsrce2`: `Claire possède un cerveau.`

   Alors, le processus de REMÉMORATION calcule la distance entre la phrase cible et chacune des phrases sources:
   la distance dis_{GN} (`phcible`, `phsrce`).
   Donc on doit calculer:
   dis_{GN} (`phcible`, `phsrce1`) = Σ_r $dis(GN_{phcible}(r), GN_{phsrce1}(r))$
   dis_{GN} (`phcible`, `phsrce2`) = Σ_r $dis(GN_{phcible}(r), GN_{phsrce1}(r))$

   dis_{GN} (`phcible`, `phsrce1`) = dis(Marie, Marie) + dis(un coeur, une poupée)
   dis_{GN} (`phcible`, `phsrce2`) = dis(Marie, Claire) + dis(un coeur, un cerveau)

4. Choisir la phrase source qui minimise la distance dis_{GN} (`phcible`, `phsrce`).
   La dis(un coeur, un cerveau) est plus petite que la dis(un coeur, une poupée) (cf.fig.2). Alors la phrase `phsrce2` est plus proche de `phcible`.

5 Conclusion

Dans ce papier, nous avons décrit un travail de recherche en cours de développement sur l'analyse de phrases en exploitant le raisonnement à partir de cas. Ce travail en est encore à ses débuts et de nombreux points restent encore à être précisés et certaines voies à explorer. Nous nous sommes jusqu'à présent intéressés à des formes grammaticales simples, mais nous envisageons pour la suite de nos travaux élargir ces principes afin de prendre en compte des formes plus variées des énoncés, des groupes nominaux plus complexes, et des phrases plus complexes. Nous envisageons pour un avenir proche de formaliser les notions proposées dans ce papier (base de cas, thésaurus, schéma de phrases, distance entre phrases, algorithmes), ainsi que les notions de chemins de similarité et d'adaptation qui ne sont qu'introduites dans ce papier. De même, nous pensons prochainement aborder le problème de la construction d'une base de cas, et celui de la construction automatique de cette base à partir d'un corpus donné. Enfin, le travail de mise en œuvre des principes de ce système sera effectué en parallèle. Toutefois, les idées à la base de ce travail nous semblent valides, cohérentes, et en tous cas prometteuses pour mener à bien un travail sur la fouille de données textuelles.

Références

[1] C.K.Riesbeck et R.C.Schank. *Inside Case-Based Reasoning*, chapitre 10, Case-Based Parsing. Lawrence Erlbaum Associates, 1989.

[2] Gregory Grefenstette. Light parsing as finite-state filtering. In *ECAI'96 workshop on "Extended finite state models of language"*, Budapest, Hungary, Aug. 11-12 1996.

[3] Jean Lieber. *Raisonnement à partir de cas et classification hiérarchique*. PhD thesis, Université Henri Poincaré - Nancy 1, 10 octobre 1997.

[4] L.Kosseim et G.Lapalme. Exibum: Un système expérimental d'extraction d'information bilingue. In *Proceedings of RIFRA-98*, Sfax, Tunisie.

[5] M.H.Al-Adhaileh et T.E.Kong. A flexible example-based parser based on the sstc. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Coling-Acl'98*, volume 1, Montréal, Québec, Canada, August 10-14 1998. Université de Montréal.

[6] M.Lenz, A.Hubner et M.Kunze. *Case-Based Reasoning Technology*, chapitre 5, Textual CBR. Springer, 1998.

[7] M.Nagao. *Artificial and Human Intelligence*, chapitre A Framework of a mechanical translation. Elsevier, 1984.

[8] E. M.Riloff. *Information Extraction as a Basis for Portable Text Classification Systems*. PhD thesis, Graduate School of the University of Massachusetts Amherst, 1994.

[9] S.Aït-Mokhtar et J.-P.Chanod. Incremental finite-state parsing. In *Proceedings of ANLP'97*, pages pp.72–79, Washington, March 31st to April 3rd 1997.

[10] S.Sato et M.Nagao. Example-based translation of technical terms. In *Proceedings of TMI-93*, pages 58–68, Koyoto, 1993.