ALSEM : Agents d'alerte sémantiques

Stéphanie WERLI (*,**)
Stephanie.Werli@paris4.sorbonne.fr

(*) LALICC – CNRS - UMR 8139, Paris IV – Sorbonne / 96, Bd Raspail – 75006 Paris, France,
(**) PSA Peugeot-Citroën, route militaire Louis Bréguet, 78140 Vélizy Villacoublay, France.

Mots clefs :

Gestion des connaissances, recherche d'information sur le Web, sémantique, extraction d'informations, base de connaissances

Keywords :

Knowledge management, information retrieval, semantics, information extraction, knowledge base

Palabras clave :

Busqueda de informacion en la red, semántica, extracción de informacion, bases de conocimiento

Résumé

Dans un contexte de maintien et de développement de la compétitivité de l'entreprise, les services de veille alimentent cette dernière en informations susceptibles d'être utilisées comme objets de référence pour l'analyse de l'environnement et l'aide à la prise de décisions.
Le projet ALSEM s'inscrit dans une démarche d'aide à la veille en déchargeant l'utilisateur d'une partie de la tâche d'exploration et de combinaison des ressources informationnelles concernant l'environnement de l'entreprise.
Ce projet, réalisé en partenariat avec l'Université Paris IV et PSA Peugeot-Citroën, vise à la conception d'un système informatique capable de collecter, d'analyser et de stocker l'information dans une base de connaissances dédiée. Il implémente notamment un procédé d'extraction de connaissances à partir des textes que nous exposons dans cet article. Ce procédé repose sur une analyse linguistique et plus particulièrement sémantique qui fournit une bonne compréhension des situations dénotées dans les documents. ALSEM implémente des techniques issues de l'ingénierie des connaissances pour l'alimentation de la base de connaissances avec les informations recueillies et la manipulation des connaissances contenues dans la base. Ces techniques autorisent la mise en place des agents d'alerte sophistiqués diffusant l'information de manière pertinente et cohérente avec le besoin du veilleur.

1 Introduction : le contexte

A l'origine, ALSEM était une extension d'un système informatique de veille sur Internet qui intégrait différents outils de recherche, de récupération d'informations sur le Web, d'indexation et d'aide à l'analyse de l'information. Ces outils, basés sur des technologies linguistiques et statistiques, avaient pour objectif de présenter aux veilleurs une vue synthétique des informations contenues dans les documents et de programmer des agents d'alerte les avertissant de l'arrivée de documents les concernant. Cependant le système a très vite démontré des limites inhérentes à tout système générique d'extraction et d'analyse d'informations à partir de sources hétérogènes (information pertinente perdue dans une masse de documents non pertinents, incapacité de capter l'information inattendue,…).
Pour améliorer les performances du système, nous nous sommes focalisés sur deux axes :

- l'amélioration du taux de pertinence des informations grâce à une délimitation précise du type d'informations recherchées et des sources d'informations privilégiées et surtout grâce à une analyse linguistique des documents et
- la mise en place de moyens de préserver l'information recueillie pour disposer d'une vision de l'environnement de l'entreprise et de son environnement.

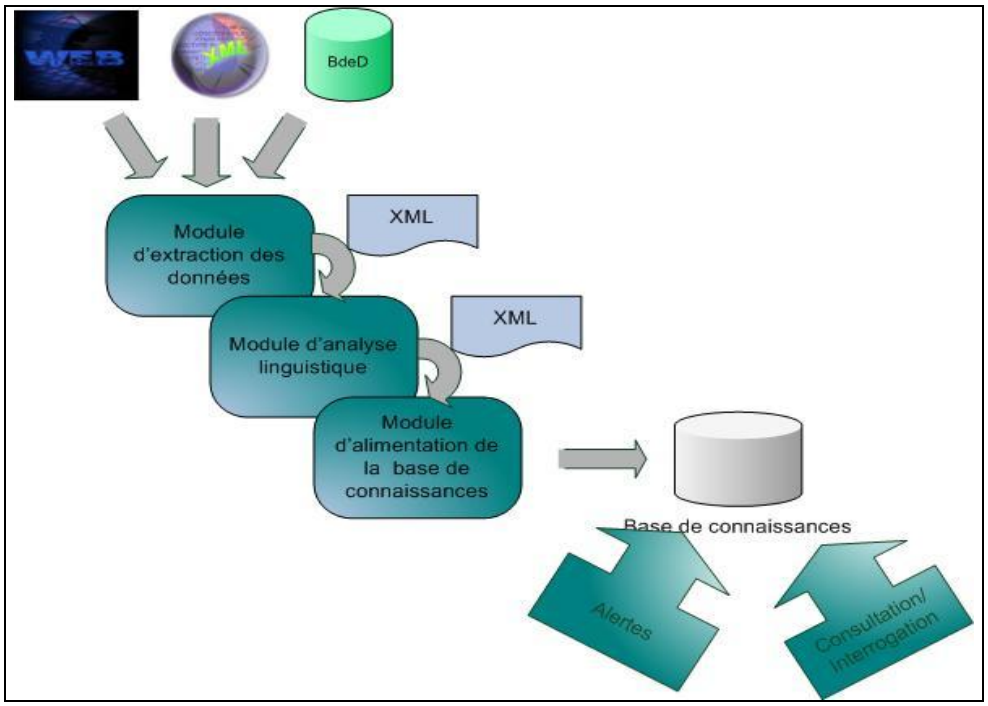
Nous avons rapidement été amenés à repenser toute l'architecture du système pour l'intégration contrôlée de l'analyse linguistique. Finalement, ALSEM a évolué en un système complet d'information intégrant divers modules dédiés aux processus de collecte, d'extraction et d'analyse d'informations à partir de données textuelles.
Dans cet article, nous exposons l'intégralité du projet de déploiement d'ALSEM.

2 Présentation du système

Le projet ALSEM propose une démarche d'ingénierie autour de l'analyse textuelle, reproductible pour différents domaines d'application et proposant des instrumentations techniques sophistiquées.
L'architecture du système s'articule autour de modules spécifiques à l'extraction d'informations à partir de textes et d'une base de connaissances organisant les informations extraites. L'architecture d'ALSEM (cf. tableau 1) intègre les modules suivants :

- un module d'extraction des données textuelles issues de sources diverses (Internet, bases de données, bases brevets, fichiers XML, fichiers textuels…),
- un module d'analyse linguistique implémentant le repérage de segments porteurs d'information pertinente dans les textes et la construction de représentations du contenu textuel,
- un module dédié à l'alimentation d'une base de connaissances avec les données et informations recueillies.

Tableau 1 : Architecture du système



Un module est constitué d'un programme principal qui prend en entrée des données et les informations nécessaires à leur analyse (sous forme de règles, de ressources, de scripts de traitement,…). L'interface de configuration associée à chaque module facilite l'édition et la combinaison de ces informations. La haute configurabilité des modules autorise leur utilisation de façon autonome ou leur intégration à d'autres applications.
Le module d'analyse linguistique et la base de connaissances forment le cœur du système, le module d'extraction des données et le module d'alimentation de la base de connaissances permettent l'intégration cohérente et contrôlée de l'analyse linguistique au sein du système et la formalisation des résultats de cette analyse. Néanmoins, ces derniers restent réutilisables dans un contexte autre que celui de l'analyse de textes.
Dans cet article, nous décrivons ces modules et leur articulation les uns par rapport aux autres à travers une étude de cas : l'analyse de dépêches et de communiqués de presse (en anglais) du secteur automobile en vue de l'extraction d'informations susceptibles d'amener des éclaircissements sur l'environnement de l'entreprise et son évolution, dans un contexte de veille stratégique.

3 L'extraction d'informations

Nous présentons, dans cette section, les stratégies mises en place pour l'extraction d'informations contenues dans les textes ou segments textuels. Notre approche pour l'accès au contenu informatif des textes s'appuie sur une analyse linguistique et plus particulièrement sémantique.
L'extraction d'informations à partir de textes passe par l'exécution des tâches suivantes : la récupération de textes ou de segments de texte susceptibles de contenir de l'information (module d'extraction des données), l'identification et le typage des objets et des relations entre objets dénotées dans les textes (module d'identification et de typage des segments porteurs d'information pertinente) et l'extraction d'une description du contenu informatif des textes (module de construction de représentations de contenu textuel).

3.1 Module d'extraction des données

3.1.1 Principe

Le module d'extraction des données permet l'extraction et la réorganisation des données structurées ou semi-structurées (pages HTML, fichiers XML, bases de données, fichiers textuels…). Au terme du

processus d'extraction des données, nous obtenons des fichiers XML (cf. tableau 1) organisant les méta-données et les données susceptibles de contenir de l'information pertinente dans un domaine d'application spécifique.

Tableau 2 : Exemple d'extraction d'une dépêche sur Internet

```
<news id="/autonews/newsId=2000.xml">
   <extraction_metadata>
      <source>AutoNews</source>
      <url>http://www.autonews.com/news.cms?newsId=2000</url>
      <extraction_date />
   </extraction_metadata >
   <headline1> GM to join China partnership by June </headline1>
   <headline2 />
   <source>Reuters</source>
   <date>May 20, 2002</date>
   <localisation>SHANGHAI</localisation>
   <body>
      <Paragraph>General Motors said on Monday it plans to set up a joint venture with
China's seventh largest automaker by June, gaining a window into what it says is a potentially
lucrative market for minivans and minitrucks.</Paragraph>
      <Paragraph>The U.S. auto giant is in the final stages of talks toward an alliance with
domestic automaker SAIC-Wuling Automobile Co. Ltd., Daphne Zheng said.</Paragraph>
   </body>
</news>
```

3.1.2 Configuration

L'extraction des données s'effectue sur la base de scripts explicitant les différentes tâches de repérage, de nettoyage, de traitement des données et spécifiant les balises XML où ces données seront stockées.
L'interface de configuration permet l'édition de ces scripts et la réutilisation de fonctionnalités implémentées dans des librairies spécifiques (par exemple, la gestion des cookies, l'extraction de chaînes de caractères contenues dans des balises HTML avec XPATH sont quelques unes des fonctionnalités intégrées à la librairie pour l'extraction de données contenues dans des pages HTML).
Le module d'extraction des données permet, par exemple, une récupération et un traitement des dépêches et des communiqués de presse adaptés au format de diffusion des sites qui fournissent un accès gratuit à ces derniers. Pour le rapatriement du corps des dépêches et des méta-données associées, un script typique explicitera les opérations suivantes :

- aller sur la page du site présentant les dernières news,
- récupérer le lien des dernières news diffusées,
- vérifier si elles n'ont pas déjà été récupérées,
- visiter les liens,
- chercher, dans les pages HTML, les balises marquant les méta-données (source, date de publication, localisation de la dépêche,…) et le corps des dépêches,
- stocker les chaînes de caractères correspondantes dans des balises XML.

3.2 Le module d'analyse textuelle

Nous considérons que l'étude du langage du domaine constitue une priorité pour une automatisation des processus de recherche, d'archivage et de traitement d'informations pertinentes. Notre système intègre l'analyse sémantique qui a pour objet la description des significations propres aux langues et leur organisation théorique.
La mise en oeuvre de stratégies d'analyse et d'extraction en adéquation avec le contenu des textes résulte d'une étude préalable sur un corpus de textes représentatifs du domaine d'application. Ces stratégies prennent donc en compte les éventuelles nuances et subtilités d'usage qui peuvent influencer

sur le sens ainsi que les caractéristiques discursives des textes (organisation, structuration…) pour l'élaboration de représentations de contenu textuel.
Ces représentations organisent les objets et faits dénotés dans l'ensemble des phrases. Elles incluent les faits importants implicites et explicites. Leurs structures s'appuient sur les spécifications des informations recherchées.
Pour construire ces représentations, notre système implémente les deux premiers niveaux d'analyse nécessaires à l'interprétation d'un texte ou d'une portion de texte : le niveau linguistique qui consiste en l'obtention d'une liste de termes et de relations directement présentes dans le texte et le niveau discursif qui explicite les relations pouvant exister entre des objets et des faits du discours.
Ces deux niveaux d'analyse sont implémentés par les deux sous-modules du module d'extraction d'information :

- le sous-module de repérage d'informations pertinentes qui identifie et type les entités nommées évoquées dans les textes (noms de compagnies, d'organisations, de lieux,…) et les formes langagières dénotant des relations pertinentes entre ces entités,
- le sous-module de construction de représentations de contenu textuel, reflets des situations essentielles dénotées dans les textes.

3.2.1 Repérage et typage de segments d'information pertinente

3.2.1.1 Principe

Il s'agit principalement de repérer et de typer les entités nommées, les groupes nominaux référant à des entités nommées et les formes langagières manifestant certaines relations entre ces entités nommées et groupes nominaux.
Ce repérage s'appuie sur analyse de surface de structures linguistiques sans le recours à l'analyse syntaxique poussée, l'information syntaxique étant incluse implicitement dans les connaissances linguistiques mises en œuvre dans l'analyse automatique.
L'analyse de texte consiste en l'appariement des séquences du texte avec des règles de correspondance. Ces règles de correspondance étendent les fonctionnalités offertes par le langage des expressions régulières. Une règle de correspondance est composée de :

- une expression régulière pouvant contenir des listes d'indices et des étiquettes sémantiques,
- une liste de conditions sur les segments identifiées par l'expression régulière,
- un ensemble de règles d'annotation sur les segments identifiés qui spécifient les informations relatives au typage des segments.

Tableau 3 : Exemple de repérage et de typage de segments d'information pertinente

```
<Sentence>
 <Company internal_id="1">General Motors</Company> <Indirect_speech type="neutral_sg" >said
<Date>on Monday</Date> it</Indirect_speech> <Intention tense="present">plans to</Agreement>
<JV_Creation tense="infinitive_present">buy out</ JV_Creation > with <Company_NG
internal_id="2">China's seventh largest automaker </Company_NG> <Temporal_Expression>by
June</ Temporal_Expression >.
</Sentence>
<Sentence>
<Company_NG internal_id="3">The U.S. auto giant <Negociation Type="final">is
in the final stages of talks</Negociation> toward <Cooperation>an
alliance</Cooperation> with </Description related_to="4">domestic automaker
.</Description> <Company internal_id="4">SAIC-Wuling Automobile Co.
Ltd.</Company>, <Source><Author><Personn last_name="Zheng">Daphne
Zheng</Personn></Author> said</Source>.
</Sentence>
```

3.2.1.2 Configuration

La configuration du processus d'identification et de typage des segments d'information pertinente s'effectue via une interface graphique. Cette interface permet d'éditer, d'importer et d'organiser des règles de correspondance, d'importer des listes d'indices,... Elle permet également de tester les règles de correspondances sur un corpus de textes.
Pour l'analyse des dépêches économiques du secteur automobile, l'interface de configuration facilite la combinaison, dans les règles de correspondance, des ressources linguistiques relatives aux trois dimensions linguistiques de ce type de texte : les termes économiques, les termes spécifiques au secteur automobile, les phénomènes linguistiques typiques du genre journalistique (le discours indirect, la source déclarative, l'insertion…).

Tableau 4 : Exemple de règle de correspondance pour le repérage d'un segment textuel dénotant la création d'une joint-venture

```
Matching_Rule => {
        Regular_Expression =>'($vb_form) &gn($determinant, $venture)',
        Conditions => [],
        Annotations=> {
                        Main_Tag => 'JV_Creation'
                        Tense => '&get_tense($1)'
                        }
}
```

Dans l'exemple ci-dessus, la règle de correspondance pour le repérage d'une thématique de création de joint-venture est composée d'une expression régulière utilisant une variable $vb_form qui contient une listes de verbes synonymes du verbe *to form* (*to create, to found, to set up*,…), une variable $venture qui contient des noms synonymes de *joint-venture* (*venture, holding company*,…) . Ces listes d'indices constituent des classes sémantique (cf. [4]). La fonction &gn(X,Y) permet d'identifier des groupes nominaux débutant par X et finissant par Y (cette identification consiste simplement à vérifier que le segment de texte correspondant ne contient pas de *stop-words* (*the, to, at, of,...*)).
Nous disposons actuellement de plus de 200 règles pour la reconnaissance des 11 thématiques économiques pertinentes dans un contexte de veille stratégique (collaborations entre entreprises concurrentes, créations de joint-venture, nomination de nouveaux dirigeants, lancement d'un nouveau produit, la fusion de deux compagnies…).

3.2.1.3 Évaluation

Pour mesurer l'efficacité de l'identification et le typage de l'information, une évaluation a été effectuée sur un nouveau corpus de dépêches extraites d'un site Web dédié au secteur automobile (différent de celui dont est issu le corpus utilisé pour l'étude linguistique). L'interprétation des résultats de l'évaluation passe par l'analyse d'indicateurs classiques :

- Rappel = réponses correctes / réponses attendues. (Cette métrique donne le taux de données pertinentes extraites par rapport à l'effectif de référence. Elle reflète la capacité d'un système à couvrir un problème),
- Précision = réponses correctes / réponses fournies. (Cette métrique reflète la qualité des réponses fournies. Un système générant autant de réponses correctes que de réponses incorrectes obtient un niveau de précision moins élevé qu'un système ne générant aucune information incorrecte).

Tableau 5 : Résultats de l'évaluation sur les entités nommées

Mesure / Entité Nommée	Rappel	Précision
Organisation	85,1%	92%
Compagnie	95,7%	98%
Personne	79%	100%
Produit	60,71%	91,9%

Tableau 6 : Résultats de l'évaluation sur les groupes nominaux référant à des entités nommées

Mesure / Groupe Nominal	Rappel	Précision
Référant à un nom d'organisation	85,1%	92%
Référant à un nom compagnie	95,7%	98%
Référant à un nom de personne	79%	100%
Référant à un nom de produit	60,71%	91,9%

Tableau 7 : Résultats de l'évaluation sur les relations

Mesure / Relation	Rappel	Précision
Accord	54,5%	100%
Processus Commerciaux	81,9%	85,5%
Négociation	26,6%	100%
Acquisition	16,7%	100%
Nomination	0%	NS
Lancement de produit	38,5%	83,3%
Investissement	60%	100%

Les tableaux 5, 6 et 7 présentent quelques résultats significatifs de l'évaluation.
Les bons résultats obtenus pour le taux de précision démontrent qu'une analyse de surface bien contrôlée est suffisante pour obtenir des résultats pertinents.
Les résultats mitigés du taux de rappel sur les relations reflètent la difficulté de couvrir toutes les diverses formulations d'un fait en prenant en compte les phénomènes de variation grammaticale ou de variation lexicale ainsi que les diverses formes de surfaces. La nomination de nouveaux dirigeants est un bon exemple de thématique difficile à couvrir (*X named Y as Z, X succeed Y as Z, X took the helm at Z, X join Y as Z,…* sont quelques unes des nombreuses façons de formuler une nomination). Néanmoins, nous continuons à capitaliser les informations linguistiques au fil de l'eau jusqu'à obtenir un taux de couverture acceptable.

3.2.2 Construction de représentations de contenu textuel

3.2.2.1 Principe

Les représentations de contenu textuel présentent de manière organisée les objets et faits dénotés dans les textes afin de refléter la situation essentielle évoquée. Dans la perspective de l'enrichissement et de la mise à jour automatiques d'une base de connaissance, une représentation doit contenir tous les attributs de la situation qui constitueront autant de point d'accès dans la base de connaissances à l'information encodée dans cette situation.

La construction des représentations passe par une analyse sémantico-conceptuelle et une réorganisation des segments d'information pertinente identifiés lors de l'étape précédente.

L'hypothèse sous-jacente au recensement des formes prédicatives est que le contexte linguistique immédiat d'un prédicat contient des informations décrivant les rôles qui y sont associés. De la structure suivante : N0 [acquire, buy out, purchase, take over, take control of, take control stake in] N1 (l'enumération des verbes forment une classe sémantique [Minel et al., 2001]), on peut déduire que N0 est l'acheteur et N1, l'acheté. La définition des rôles occupés par les constituants est déductible des structures syntaxiques dans lesquelles ils occurrent. Un système comme Autoslog [Riloff, 1994] est basé sur cette hypothèse.

Une fois que les rôles ont été identifiés et reliés aux prédicats, le système réorganise les segments porteurs d'information pertinente sous la forme de formulaires à champs prédéfinis (selon la thématique, i.e. le prédicat identifié) où les valeurs des champs sont des chaînes de caractères.

Tableau 8 : Exemples de représentation de contenu phrastique

```
<Analysis textID="/autonews/newsId=2000.xml">
<RELATES paragraph="1" sentence="1">
  <JV_Creation>
    <ACTORS>
      <Company>
        <NAME>General Motors</NAME>
      </Company>
      <Company>
        <DESCRIPTION>China's seventh largest automaker
          </DESCRIPTION>
      </Company>
    </ACTORS>
    <DATE>by June</DATE>
  </JV_Creation>
</RELATES>


</Analysis>
<Analysis textID="/autonews/newsId=2000.xml">
<RELATES paragraph="2" sentence="2">
  <Cooperation>
    <ACTORS>
      <Company>
        <DESCRIPTION>The U.S. auto
          giant</DESCRIPTION>
      </Company>
      <Company>
        <NAME>SAIC-Wuling Automobile Co. Ltd.</NAME>
      </Company>
    </ACTORS>
    <STATUS>negociation</STATUS>
    <SOURCE>
      <PERSON>Daphne Zheng</PERSON>
```

```
            </SOURCE>
        </Cooperation>
    </RELATES>
</Analysis>
```

Tableau 9 : Exemple de représentation de contenu textuel

```
<Analysis textID="/autonews/newsId=2000.xml">
<RELATES paragraph="1" sentence="1">
    <JV_Creation>
        <ACTORS>
            <Company>
                <NAME>General Motors</NAME>
                <DESCRIPTION>The U.S. auto giant
                    </DESCRIPTION>
            </Company>
            <Company>
                <NAME> SAIC-Wuling Automobile Co. Ltd.</NAME>
                <DESCRIPTION> China's seventh largest automaker
                    </DESCRIPTION>
            </Company>
        </ACTORS>
        <DATE>by June</DATE>
    <STATUS>negociation</STATUS>
    <SOURCE>
        <PERSON>Daphne Zheng</PERSON>
    </SOURCE>
    </JV_Creation>
</RELATES>
</Analysis>
```

3.2.2.2 Configuration

La configuration du module de construction de représentation consiste en l'édition de règles de réorganisation et de reformulation des segments porteurs d'information pertinente afin d'obtenir une représentation de l'information saillante relatée au niveau de la phrase.

La difficulté majeure rencontrée lors de la construction de représentation au niveau phrastique concerne la capture de relations implicites spécifiques au domaine. Les relations implicites sont des relations déductibles par des inférences spécifiques au domaine d'application (par exemple, de l'analyse sémantique de *X has signed an agreement for Y developed by Z*, les deux informations suivantes sont déduites : X coopère avec Y et Y développe le produit Y. De ces deux informations adjacentes, le module de construction de représentation infère que X coopère avec Y sur un produit Z).

Nous ne maîtrisons pas encore complètement la construction de représentations au niveau textuel qui nécessite d'implémenter des solutions complexes pour la fusion des représentations obtenues au niveau phrastique en résolvant la coréférence d'entités (la coréférence associe deux syntagmes nominaux référant à une même entité) et la coréférence événementielle (les conférences MUC-6 et MUC-7 [5] sont riches d'enseignements sur la problématique du calcul de la chaîne de coréférence dans les textes).

Avant de continuer sur le module d'alimentation de la base de connaissances, nous présentons brièvement la structuration de la base de connaissances.

4 Stockage d'informations

Afin de capter efficacement les évolutions de son environnement, le veilleur doit pouvoir accéder à une description opérationnelle de cet environnement économique adaptée aux ambitions propres à l'entreprise. Notre système exploite donc une base de connaissances décrivant les aspects pertinents de l'environnement de l'entreprise.

4.1 Présentation de la base de connaissances

4.1.1 Modélisation

Pour matérialiser la base de connaissances, nous nous sommes orientés dans un premier temps vers le développement d'une ontologie. Une ontologie fournit les ressources conceptuelles et notionnelles pour formuler et expliciter de manière systématique les notions utiles à la formulation des connaissances [2].
Notre ontologie définit donc les primitives nécessaires à la représentation de l'information extraite. Pour sa conception, nous avons utilisé Protégé [6], environnement graphique de développement d'ontologies développé par le SMI de Stanford. Protégé fournit une interface graphique pour la modélisation des classes (concepts du domaine), leurs attributs et les relations entre concepts et divers outils d'exploitation de la base (visualisation graphique, moteur d'inférence, lancement de requêtes…).

4.1.2 Structuration de la base de connaissances

Actuellement, notre base de connaissances se décline en trois sous bases :

- une base des entités nommées contenant des informations basiques sur les acteurs de l'environnement du type nom, alias, localisation,… (ces informations sont issues de base de données ou de sites donnant des informations générales sur les entreprises),
- une base des événements contenant les différents aspects relatifs aux événements comme les acteurs intervenant dans l'événement, le statut temporel (passé, futur, en cours) de l'événement ou sa date d'occurrence, la source de l'information, l'identifiant du texte où l'événement a été relaté,…
- une base documentaire contenant les textes collectés et les méta-données.

4.1.3 Le Module d'alimentation de la base de connaissances

4.1.3.1 Principe

Le module d'alimentation répond à deux problématiques : la transformation des objets et faits des représentations en objets de la base de connaissances et l'appariement des nouvelles informations avec celles contenues dans la base.

4.1.3.2 Configuration

La configuration du module d'alimentation de la base de connaissances prend en charge l'instanciation des objets dans la base à partir de l'ordre et de la hiérarchie des balises XML des représentations de contenu textuel. L'implémentation de la base repose donc sur les choix de normalisation sémantique résultant du travail sur le corpus de textes représentatifs du domaine.
L'interface de configuration gère la spécification des vérifications nécessaires lors de la création ou de la mise à jour d'un objet selon son type. Les règles implémentées jusqu'à présent restent basiques et n'exploitent pas encore réellement les outils d'inférence à disposition. Une règle simple pour l'instanciation d'un objet de la classe *Company* consistera, par exemple, à vérifier que l'objet n'est pas existant en comparant les valeurs des propriétés *Name* et *Alias*.

L'interface de configuration permet également d'intégrer certaines fonctionnalités pour la transformation des chaînes de caractères en données assimilables par le système (comme le calcul de la date).

4.2 Perspectives autour de la base de connaissances

Notre première ébauche de base de connaissances présente des résultats satisfaisants et compréhensibles pour l'utilisateur.
Cependant, une ontologie se révèle relativement inadaptée dans un contexte de création de sens (au sens de Lesca [3]) sur la base d'informations disparates, parfois incomplètes, contradictoires ou ambiguës et dans un contexte d'accès à l'information en fonction d'un contexte donné.
Nous nous orientons donc aujourd'hui vers la conception d'une base de connaissances représentant l'environnement économique de l'entreprise et capable d'évoluer en adéquation avec l'occurrence d'événements significatifs. Pour sa conception, nous nous concentrerons sur les aspects de la conceptualisation nécessaires au bon fonctionnement du système et aux fonctionnalités étendues de navigation et d'interrogation (comme, par exemple, la possibilité de retrouver à partir d'un objet de la base tout ou une partie de l'information qui lui est connexe).

4.2.1 Vers une base de donnée relationnelle

Pour l'implémentation de cette nouvelle base de connaissance, nous explorons la piste des bases de données de type relationnelle où les relations sont des objets de la base au même titre que les entités.
L'implémentation d'une base de connaissances basée sur la méthode AOM (*Asset-Oriented Modeling*) [1] semble ici une bonne alternative.
En effet, AOM autorise la réification des relations et donc la définition de propriétés sur les relations et de relations entre les relations. De plus, AOM permet de construire plusieurs niveaux de représentation d'un objet et donc d'accéder à des niveaux de granularité différents.

4.2.2 Agents d'alerte

La principale fonctionnalité développée autour de la base de connaissances concerne la spécification d'agents d'alerte capables d'identifier des changements significatifs dans l'environnement de l'entreprise selon des thématiques spécifiques (le rapprochement de deux sociétés, l'émergence d'une nouvelle technologie, d'un nouveau marché, de nouvelles tendances, de nouveaux acteurs,...).
L'ontologie autorise la mise en place d'agents d'alerte sur l'occurrence d'événements intéressants.
Nous étudions également, en parallèle, l'apport d'une base de données relationnelle pour la mise en place d'agents d'alerte plus sophistiqués capables de combiner une nouvelle information, un nouvel événement avec les informations contenues dans la base. De plus, l'exploitation d'une base relationnelle permettrait de donner la possibilité à l'utilisateur final de spécifier des alertes sur n'importe quel objet ou type d'objets de la base (et donc d'être alerter sur les changements de propriétés d'objets, les créations ou destructions d'objets,…)

5 Conclusion

Nous disposons aujourd'hui d'un dispositif évolutif, organisé autour de l'analyse linguistique et développé avec des technologies souples autorisant l'intégration ultérieure de fonctionnalités ou de procédés qui restent pour l'instant difficiles à implémenter (comme l'analyse de la coréférence, cf. 3.3.3.2).
L'expérience acquise nous permet aujourd'hui de mieux appréhender la problématique de l'analyse linguistique pour l'extraction d'informations à partir de textes. Il nous reste aujourd'hui à vérifier que notre approche est facilement transposable à d'autres domaines et à d'autres types de textes (brevets, courrier clients,...).

6 Bibliographie

[1] AOM, htpp://www.aomodeling.org/

[2] BACHIMONT B., *Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle*, IC 2001.

[3] LESCA H., 2002, *Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles*, Colloque CIFPME 2002, Montréal, Québec, Canada.

[4] MINEL J-L, DESCLES J-P, CARTIER E., CRISPINO G., BEN HAZEZ S., JACKIEWICZ A., *Résumé automatique par filtrage sémantique d'informations dans des textes*, Technique et Science Informatiques, 2001.

[5] MUC, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[6] PROTÉGÉ-2000, http://protege.stanford.edu

[7] RILOFF E., *Information Extraction as a Basis for Portable Text Classification Systems*, Thèse de doctorat, Université du Massachussets, Amherst.

[8] WERLI S., *Agents de détection d'événements et de faits majeurs : Apllication à la veille*, Thèse de doctorat, en cours

Remerciements

Je remercie Jean-Luc Minel pour la relecture attentive de cet article.