Le Text Mining multilingue : application au monde de l'intelligence économique[1]

Pascal Coupet (*), Bianka Buschbeck (*), Amandine Six (*), Françoise Cardoso (*) et Charles Huot (*)
Pascal.Coupet@temis-group.com, Bianka.Buschbeck@temis-group.com, Amandine.Six@temis-group.com, Francoise.Cardoso@temis-group.com, Charles.Huot@temis-group.com

(*) : TEMIS SA, Tour Gamma, 193-197 rue de Bercy, 75582 Paris cedex 12, www.temis.com .

Mots clés : text mining, intelligence économique, analyse automatique de texte, multilinguisme.

Résumé

Le Text Mining a démontré depuis quelques années son efficacité dans la génération de valeur ajoutée lors de l'analyse automatique de flux de presse économique. Dans la majorité des cas, ces flux de presse sont monolingues. Les règles d'extraction d'information ainsi que les lexiques utilisés sont le plus souvent dépendants de la langue que l'on souhaite analyser.

Nous présenterons dans cet article les réalisations de la société TEMIS, ainsi que le modèle que nous avons développé et mis en place afin d'analyser de façon homogène des flux de presse économique en 5 langues. Nous présenterons notamment la technologie dite des Cartouches de Connaissances ™ ou Skill Cartridges™ qui permet ce nouveau type d'extraction multilingue. A titre d'exemple, nous présenterons les résultats obtenus sur l'analyse d'un corpus en Français, Anglais, Espagnol, Italien et Allemand.

[1] Marques déposées : Insight Discoverer, Cartouches de connaissances, Skill Cartridges et Online Miner et Xelda sont des marques déposées par TEMIS SA en France et dans le reste du monde.

1. Introduction

Alors que la langue anglaise prédomine dans la littérature et les bases de données d'information scientifique et technique, la situation est différente pour l'information presse, économique ou juridique. Pour ces dernières, chaque pays est producteur d'information dans sa propre langue. Ainsi depuis les années soixante dix et le développement massif des bases de données électroniques, les méthodes et techniques de traitement de l'information scientifique pour réaliser une veille technologique efficace ont la plus part du temps été confrontées soit à une information semi structurée de type champ de descripteurs codés [1], [2] soit à l'analyse de résumés en anglais [3].

Les sociétés ou les états souhaitant mettre en place des systèmes de surveillance et des moyens d'analyse couvrant les secteurs qui sont au-delà de la science et de la technologie ont éprouvé naturellement le besoin d'outils de traitement de l'information pour une grande variété de langues [4]. Nous présentons dans cet article l'un de ces logiciels développé par la société TEMIS et mis en production dans diverses grandes sociétés européennes. L'originalité de notre approche repose sur l'utilisation combinée de méthodes d'analyse morphosyntaxique et sémantique à travers un concept unique : La cartouche de connaissance ou Skill Cartridge. Ceci permet d'analyser avec la même logique applicative des documents dans des langues différentes en restituant les résultats sous la forme de concepts génériques. Nous donnerons une explication sur l'approche linguistique utilisée à travers la description du moteur Xelda, analyseur morphosyntaxique multilingue, puis du système d'extraction, Insight Discoverer Extractor et enfin nous décrirons l'architecture de la Cartouche de Connaissance multilingue Competitive Intelligence.

Afin d'illustrer ces différents aspects de l'analyse linguistique, nous fournirons des exemples tout au long de l'article, à travers l'explication de copies d'écran de l'application.

En conclusion nous aborderons le futur de l'approche présentée ici à travers ces extensions possible vers de nouvelles applications.

2. Les étapes de l'analyse linguistique

2.1. Le schéma général de fonctionnement

Les flux d'information économique sont la matière première des systèmes d'intelligence économique. Pour leur exploitation efficace, les logiciels de text mining vont en faire une analyse en profondeur et générer des rapports de synthèses destinés à des analystes et dans nombre de cas au lecteur final [5].

La première étape consiste, après la phase d'acquisition du contenu, à lire un premier document. Son format peut être variable (fichier de type *pdf, doc, ppt, zip*, etc…). Il est converti dans un format de type *xml*.

La seconde étape consiste à identifier la langue du document afin de lui appliquer le bon module d'analyse.

La troisième étape est l'analyse linguistique profonde réalisée par l'analyseur morphosyntaxique Xelda.

La quatrième étape est l'extraction des concepts et des relations entre les concepts.

La cinquième étape consiste à stocker cette information dans une base de données.

Enfin la sixième étape intervient à la fin du traitement du dernier document. C'est la phase de génération de rapport de synthèse. Dans cette phase l'ensemble des concepts et des relations extraites est agrégé dans une interface unique de navigation. L'utilisateur peut alors naviguer par les concepts ou par les relations à travers l'ensemble des documents et ceci indépendamment de la langue d'origine de ces derniers.

2.2. L'analyseur morphosyntaxique Xelda

L'analyseur morphosyntaxique Xelda utilise la technologie linguistique XFST (Technologie des automates à états finis développé par Xerox). Il offre une gamme évolutive de services fondés sur des composants de traitement du langage naturel qui s'intègre dans des applications d'entreprise. Parmi les fonctionnalités les plus importantes dans le processus d'analyse de document, nous trouvons les suivantes :

-*identification de la langue* : reconnaît automatiquement la langue utilisée dans chaque document,
-*segmentation* : divise un texte en phrases,
-*tokenisation* : scinde un texte en unités lexicales élémentaires,
-*analyse morphologique* : renvoie les catégories grammaticales potentielles pour tous les mots identifiés durant la phase de tokenisation,
-*désambiguïsation morpho-syntaxique* : détermine la catégorie grammaticale d'un mot en fonction de son contexte.

Xelda peut analyser des documents dans 16 langues :
(-)allemand, anglais, espagnol, français, grec, hongrois, italien, néerlandais, polonais, portugais, russe, tchèque, norvégien, finlandais, suédois et danois. TEMIS travaille actuellement sur plusieurs langues supplémentaires notamment comme l'Arabe.

2.3. L'extracteur Insight Discoverer Extractor (IDE)

La phase d'extraction d'information multilingue proprement dite est réalisée par le moteur Insight Discoverer Extractor. Il s'agit d'un serveur d'extraction d'information dédié à l'analyse de documents textuels. Il enchaîne trois étapes:

-*l'analyse de corpus* : lecture de 50 formats de fichiers incluant l'identification automatique de la langue,
-*l'analyse linguistique* réalisée par Xelda
-*l'extraction de connaissance* (exécution des règles d'extraction) :
    -identification des entités (noms de personnes, noms de compagnies, noms des produits, dates, lieux, etc…) pour les applications en Intelligence économique
    -reconnaissance des relations entre les entités (relations d'achat, de cause à effet entre 2 sociétés ou deux pays, etc…)
-*Normalisation et validation des concepts identifiés*

L'extraction d'information est le processus qui permet d'identifier l'information pertinente, les critères de pertinence étant exprimés sous forme de composants de connaissance rassemblés dans une cartouche de connaissance.
IDE utilise comme Xelda la technologie des automates à états finis (transducteurs).
Parmi les langues analysées par IDE, on retrouve le français, l'anglais, l'espagnol, l'italien et l'allemand.
Ainsi la richesse d'analyse automatique que l'on obtient par l'extraction d'information va être guidée par la cartouche de connaissance.

2.4. La Cartouche de Connaissance Intelligence Economique

2.4.1. Le concept de Cartouche de Connaissance

Il s'agit d'un concept développé par la société TEMIS depuis la fin des années 1990 (2000)[6]. L'idée originale a vu le jour à la suite de nombreux projets de Text Mining réalisés auprès de clients dans le monde entier. Afin de préserver l'expérience acquise dans l'extraction de concepts complexes, il était nécessaire de créer un récipient dans lequel il soit possible de la stocker en vue d'une réutilisation future. Ceci devant permettre une réduction

drastique de la déperdition d'énergie constatée projet après projet. Puisque le moteur d'extraction restait stable au cours du temps les éléments variables de pilotage de cette extraction devaient être stockés à la périphérie du système. Le concept d'avoir des cartouches venant selon les besoins se greffer sur l'IDE a permis une très grande modularité nécessaire pour la réalisation de nombreuses applications. Aujourd'hui TEMIS dispose d'une librairie de cartouches de connaissances couvrant des secteurs comme l'intelligence économique, l'analyse de la relation client, la médecine, la biologie [7], la chimie, les ressources humaines, les entités génériques (noms, lieux, dates, mesures,etc…), le juridique, et ceci dans une grande variété de langues. Chaque cartouche est spécifiée, développée, documentée, maintenue et fait l'objet d'un plan qualité au même titre qu'un autre logiciel de la gamme. Des versions majeures et mineures sont livrées aux clients de manière régulière.

2.4.2. La Cartouche de Connaissance Intelligence Economique (Competitive Intelligence CI)

Parmi les premiers secteurs ayant fait l'objet d'un effort très important de développement, se trouve l'intelligence économique [8] (Competitive Intelligence). Dès le début des années 2001, nous avons travaillé aux premières versions de cette cartouche. Aujourd'hui nous en sommes à la version 2.0 multilingue.

2.4.2.1. Objectif de la cartouche de connaissance Competitive Intelligence (CI)

La cartouche CI est un produit qui permet d'identifier un ensemble de relations prédéfinies entre tous les acteurs d'une industrie donnée. Elle fournit une information précise comme les résultats financiers d'une entreprise, son environnement concurrentiel, les produits du marché ainsi que les stratégies de développement des acteurs.

A titre d'exemple cette cartouche permet de répondre aux questions suivantes :

-qui fusionne avec qui ?
-quel est le montant investi dans une fusion ?
-cette fusion est elle effective ou s'agit il d'une annonce ?
- quel est le chiffre d'affaire d'une société donnée ?
-quelle entreprise investit dans quel secteur ?
-qui est intéressé dans le rachat d'une entreprise donnée ?

2.4.2.2. Structure de la cartouche

La cartouche est basée sur une architecture de composants. Ceci permet une indépendance entre chaque composant ou *unité de cartouche*. Cette structure permet également de faciliter la maintenance de la cartouche.
La cartouche CI est composée d'un ensemble d'unité basique et d'unité thématique.

2.4.2.2.1. Le composant Competitive Intelligence Basic

Il existe trois unités basiques regroupées sous l'intitulé Competitive Intelligence basic components :

-les tools

C'est un composant de reconnaissance général qui identifie les termes généraux et les expressions fréquentes. Il est multilingue et possède la même couverture sémantique pour l'anglais, le français, l'espagnol, l'italien, l'allemand, le portugais et le néerlandais. Les concepts identifiés sont les suivants :

-expression monétaire

-nombres
-unités de date et de temps
-distances
-poids et mesures
-adresses électroniques

-la reconnaissance d'entités nommées

C'est un composant de reconnaissance général qui identifie les noms propres. Il combine des lexiques spécifiques par domaine ainsi que des règles de détection automatique en contexte. Il possède une couverture de langues équivalente au composant *tools*. Les concepts identifiés sont les suivants :

-Noms de personnes
-Lieux
-Noms de sociétés
-Noms d'organisations
-Noms d'établissements
-Noms de produits

-les dictionnaires

Ce composant contient des lexiques de terminologie générale couvrant un ensemble de champs sémantiques. Ce composant contient également un ensemble de lexiques spécifiques à certains domaines comme par exemple la terminologie biomédicale.

Au niveau du composant Competitive Intelligence basic, il est possible de réaliser un certain nombre de tâches comme par exemple combiner un concept *expression monétaire* et un concept *nombres*, pour obtenir un *montant financier*.

2.4.2.2.2. Le composant Competitive Intelligence Thématique

Ce composant contient un ensemble de concepts d'intelligence économique. Chacun de ces concepts correspond à un champ sémantique particulier de l'intelligence économique. La couverture est la même pour l'anglais, le français, l'espagnol, l'allemand et l'italien. Les thèmes et sous thèmes couverts sont les suivants :

-Information financière
-reporting financier
-chiffre d'affaire
-vente
-notation d'agences

-Intelligence d'entreprise
-cession(s) et acquisition(s)
-fusion
-OPA
-Stratégie
-partenariat de co-développement
-partenariat de co-marketing
-licensing
-Actif
-propriété
-prise de participation
-actionnariat
-levée de fonds
-capital

-privatisation
-Management
-fonction dans l'entreprise
-changement de fonction
-Investissement
-investissement
-expansion
-Marché et produits
-part de marché
-lancement de produit
-vente de produit
-Restructuration
-désinvestissement
-faillite

La Cartouche de Connaissance Intelligence Economique

2.4.2.3. Les rôles dans la cartouche d'intelligence économique

Les relations sémantiques fournies par la cartouche CI font référence à des entités nommées comme à des patrons spécifiques d'intelligence économique. Chaque entité se voit assignée un rôle qui définit sa position dans la relation. Les rôles utilisés dans la cartouche CI sont les suivants :

Rôle	**Description**
Who	Fait référence à l'acteur qui est le sujet de la relation (société, organisation, personne…)
Whom	Fait référence à l'acteur qui est l'objet de la relation. Ce rôle peut avoir diverses instantiations telles que : to whom, from whom, of whom.
What	Fait référence au produit, à la technologie, à un service, à une division, à une filiale…
Where	Fait référence au lieu de l'action
When	Fait référence au temps Exemples : l'année précédente, le premier trimestre 2005, l'année prochaine…
CI pattern	Fait référence au thème générique d'intelligence économique Exemples : a investi, fait un partenariat,…
Rumor	Fait référence à des expressions qui indiquent la probabilité d'un événement ou d'une action Exemples : peut faire l'acquisition, envisage de, est intéressé à, prévoit,…
Annoucement	Fait référence à des mots ou des expressions qui introduisent les événements Exemples : annonce, publie, indique,…

En plus des rôles, la cartouche recense les expressions sémantiques utilisées par la plupart des relations.

Expression	**Description**
Fluctuation	Fait référence à des variations pour les montants monétaires comme pour l'activité. Les indications comme la dépréciation ou l'évaluation sont intégrées dans *Fluctuation*. Exemples : décroître, réduire, décliner, diminuer, ralentir, stagner, évaluer, augmenter, doubler, surclasser, déclasser…
Money amount	Fait référence à des montants financiers et des pourcentages de croissance de chiffres d'affaires Exemple : € 150 million, +9,5%...

3. Les applications à l'intelligence économique

Comme nous l'avons décrit dans la section schéma général de fonctionnement, le résultat de l'analyse d'un flux d'information presse est finalement agrégé dans un rapport de synthèse.

Ce rapport permet une navigation au sein des concepts extraits. Une illustration de l'interface est donnée dans la copie d'écran qui suit (figure 1).

Comme on peut le voir dans le cadre situé en bas à gauche, les concepts détectés dans ce cas sont à la fois en anglais et en français.

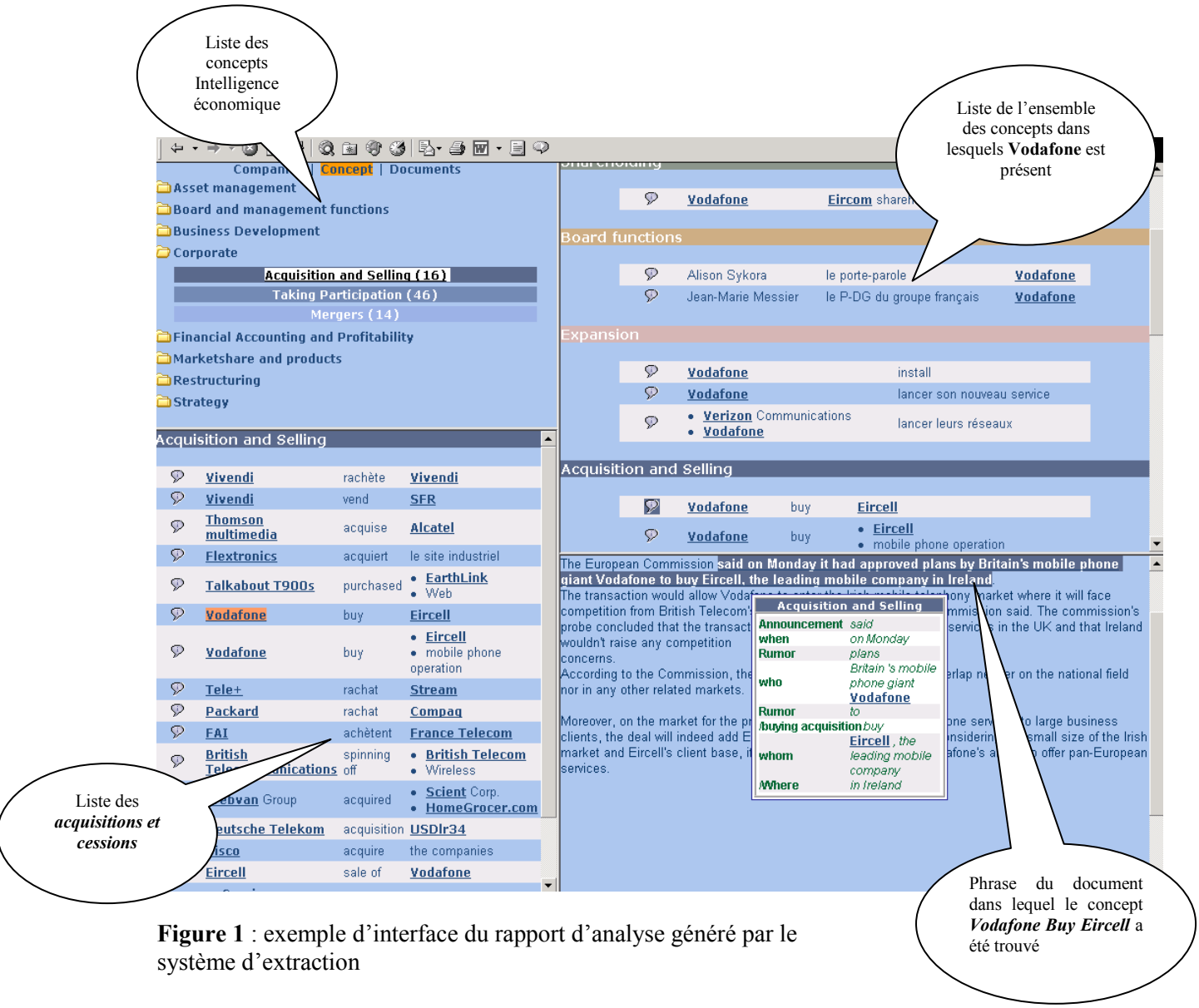

Figure 1 : exemple d'interface du rapport d'analyse généré par le système d'extraction

Cette table récapitule les divers rôles extraits dans la phrase.

Acquisition and Selling	
Announcement	said
when	on Monday
Rumor	plans
who	Britain 's mobile phone giant Vodafone
Rumor	to
/buying acquisition	buy
whom	Eircell , the leading mobile company
Where	in Ireland

L'exemple ci-dessous nous montre le même concept détecté en espagnol.

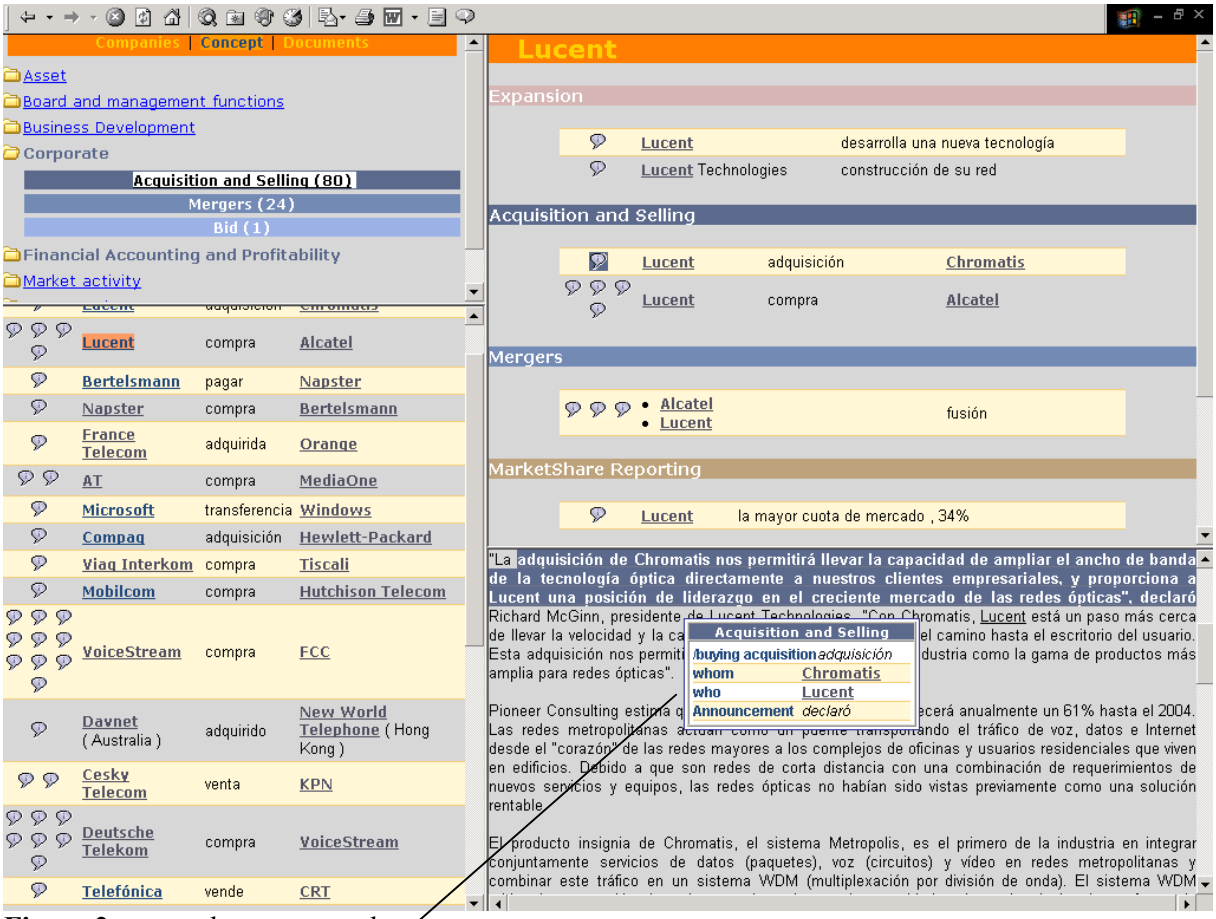

Figure 2 : exemple en espagnol

Acquisition and Selling	
/buying acquisition	adquisición
whom	Chromatis
who	Lucent
Announcement	declaró

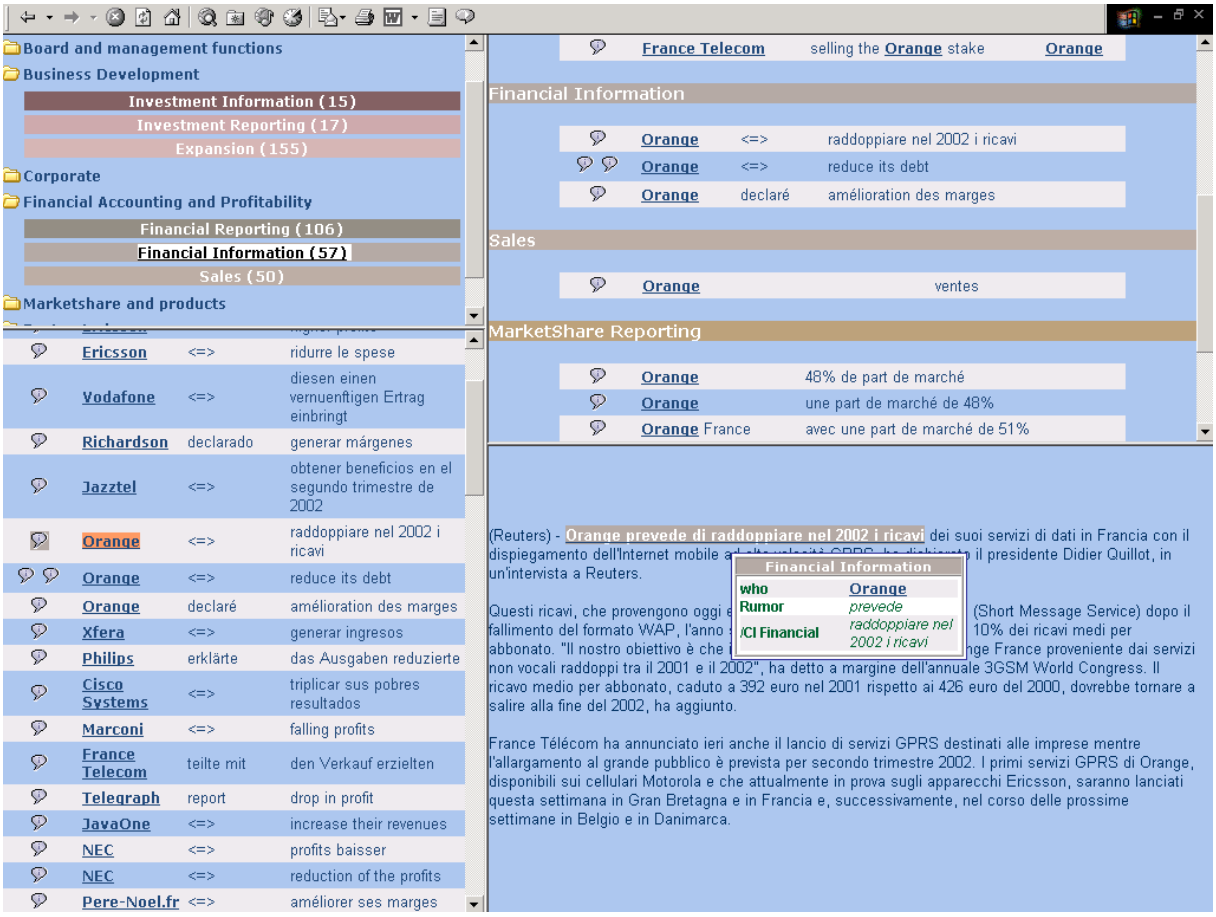
Board and management functions
Business Development
Investment Information (15)
Investment Reporting (17)
Expansion (155)
Corporate
Financial Accounting and Profitability
Financial Reporting (106)
Financial Information (57)
Sales (50)
Marketshare and products
Ericsson <=> ridurre le spese
Vodafone <=> diesen einen vernuenftigen Ertrag einbringt
Richardson declarado generar márgenes
Jazztel <=> obtener beneficios en el segundo trimestre de 2002
Orange <=> raddoppiare nel 2002 i ricavi
Orange <=> reduce its debt
Orange declaré amélioration des marges
Xfera <=> generar ingresos
Philips erklärte das Ausgaben reduzierte
Cisco Systems <=> triplicar sus pobres resultados
Marconi <=> falling profits
France Telecom teilte mit den Verkauf erzielten
Telegraph report drop in profit
JavaOne <=> increase their revenues
NEC <=> profits baisser
NEC <=> reduction of the profits
Pere-Noel.fr <=> améliorer ses marges
France Telecom selling the Orange stake Orange
Financial Information
Orange <=> raddoppiare nel 2002 i ricavi
Orange <=> reduce its debt
Orange declaré amélioration des marges
Sales
Orange ventes
MarketShare Reporting
Orange 48% de part de marché
Orange une part de marché de 48%
Orange France avec une part de marché de 51%
(Reuters) - Orange prevede di raddoppiare nel 2002 i ricav dei suoi servizi di dati in Francia con il dispiegamento dell'Internet mobile il presidente Didier Quillot, in un'intervista a Reuters.
Financial Information
who Orange
Rumor prevede
/CI Financial raddoppiare nel 2002 i ricavi
Questi ricavi, che provengono oggi (Short Message Service) dopo il fallimento del formato WAP, l'anno 10% dei ricavi medi per abbonato. "Il nostro obiettivo è che nge France proveniente dai servizi non vocali raddoppi tra il 2001 e il 2002", ha detto a margine dell'annuale 3GSM World Congress. Il ricavo medio per abbonato, caduto a 392 euro nel 2001 rispetto ai 426 euro del 2000, dovrebbe tornare a salire alla fine del 2002, ha aggiunto.
France Télécom ha annunciato ieri anche il lancio di servizi GPRS destinati alle imprese mentre l'allargamento al grande pubblico è prevista per secondo trimestre 2002. I primi servizi GPRS di Orange, disponibili sui cellulari Motorola e che attualmente in prova sugli apparecchi Ericsson, saranno lanciati questa settimana in Gran Bretagna e in Francia e, successivamente, nel corso delle prossime settimane in Belgio e in Danimarca.

Figure 4

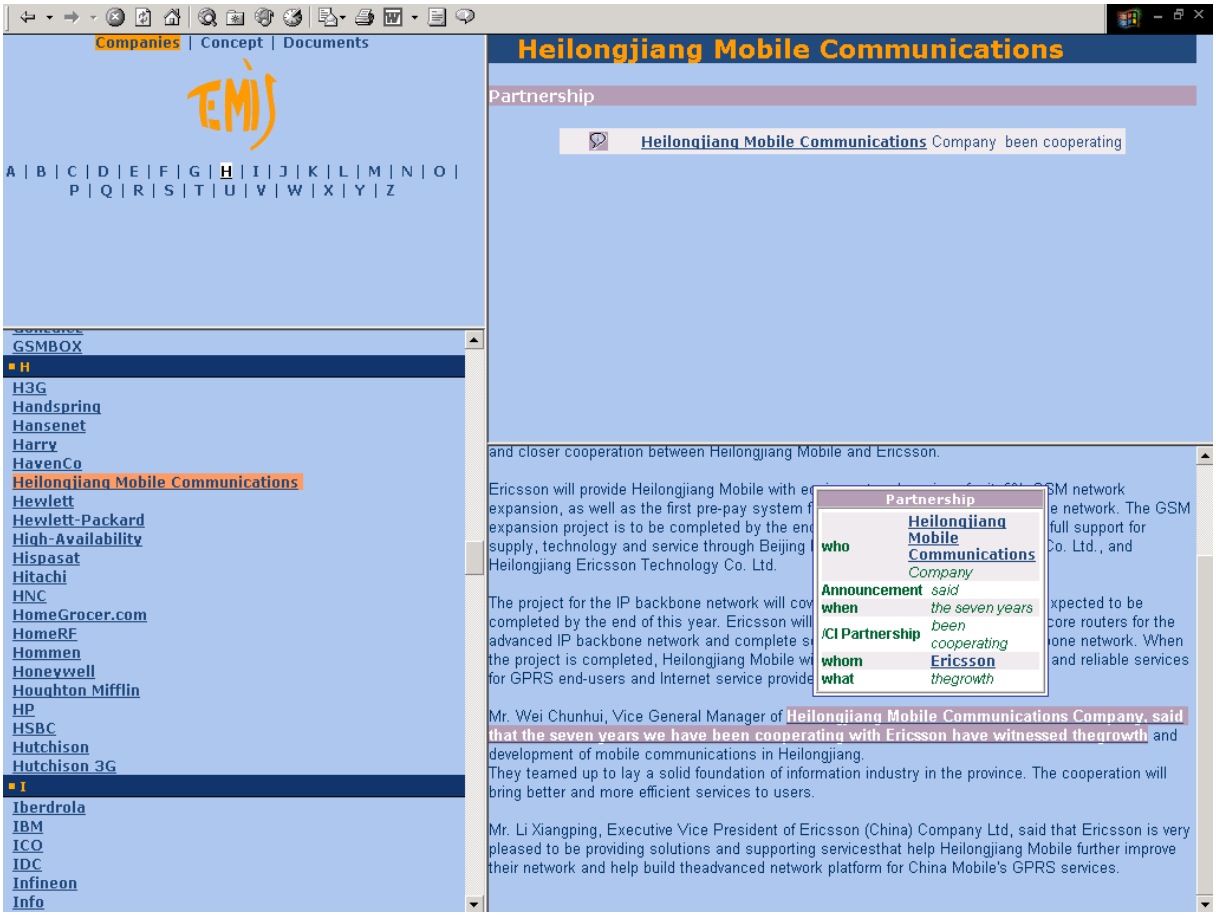

Figure 5 : Exemple de navigation dans le rapport via la liste des sociétés identifiées par le système.

4. Conclusion et perspectives

Ces travaux ont montré l'intérêt de l'analyse linguiste multilingue et de l'extraction d'information automatisée dans l'aide qu'elle apporte aux entreprises pour leur activité d'intelligence économique. L'apport du concept de cartouche de connaissance comme réceptacle de la terminologie applicative et des règles métiers permet une meilleure gestion et maintenance des éléments de connaissance du veilleur.

Les perspectives de développement se structurent selon trois axes aujourd'hui.

Le premier axe concerne les applications. Au-delà de l'intelligence économique, d'autres domaines applicatifs ont le besoin d'analyser des textes dans des langues diverses mais aussi de présenter une synthèse unique. C'est le cas notamment de l'analyse de la relation clientèle ou de l'analyse de l'évolution des besoins, des envies ou des goûts des consommateurs. Une grande quantité d'information textuelle est disponible dans ce domaine, information recueillie dans les centres d'appel, les sites internet(s), les forums, etc…Le secteur des ressources humaines dans les sociétés multinationales doit également faire face à une multitude de textes multilingues (CV, lettre de motivation, bases de compétences etc…).

Le second axe concerne les langues. Pour chacune des applications, le développement des cartouches dans des langues supplémentaires est nécessaire suivant le potentiel marché disponible.

Enfin chaque secteur industriel possède ces propres spécificités qui se doivent d'être modélisées dans des cartouches. Nous avons ainsi développé une cartouche d'intelligence économique spécifique à l'industrie pharmaceutique (la cartouche de connaissance Competitive Intelligence Life Science Edition) qui reprend les concepts standard de la cartouche CI en lui ajoutant des concepts spécifiques

comme « autorisation de mise sur le marché », « aire thérapeutique », ou « essais cliniques ». Nous poursuivons également le développement de cartouche dans les secteurs de l'énergie, de l'édition, de l'automobile, de l'agroalimentaire ou de la chimie.

Comme nous le voyons, les axes de développement ne manquent pas. C'est pourquoi nous avons mis en place un programme de collaboration avec des universités et des instituts de recherche principalement en Europe. L'objet de ces collaborations est le développement, la mise à jour et la maintenance de cartouches de connaissances dans le secteur de compétence de nos partenaires. Ce programme permet à chacun des partenaires de se focaliser sur la modélisation d'une partie de sa connaissance en utilisant des outils développés par TEMIS comme le *Skill Cartridge Studio*.

5. Bibliographie

[1] : Rostaing H., "Traitement de l'information à l'aide d'outils infométriques " (1999), Dans le cadre du Réseau Européen REVEIL , organisé par le Centre de Recherche Publique Henri Tudor, Luxembourg, le 6 mai , (1999)

[2] : C .Bédécarrax and C.Huot (1993). A new methodology for systematic exploitation of technology databases. In Information Processing & Management Vol. 30, No. 2, 1993, Pergamon Press Ltd.

[3] : Massetani M., Neri F., Priamo A., Relevant terminologies as descriptors for documents (2002), SYNTHEMA Lexical Systems Lab, internal report, Pisa, Italy, pp. 1-7, 2002.

[4] : B.DELECROIX, S.GUILLEMIN-LANNE, A.SIX (2004).Veille concurrentielle et Veille stratégique: deux applications d'extraction d'information, Toulouse, 25-29 octobre, 2004, FPC/UPC-SFBA-IRIT.

[5] : B.Buschbeck, L.Grivel, S.Guillemin-Lanne, C.Lautier (2001), Une application industrielle d'extraction de l'information pour l'intelligence économique. ECG 2002 Actes Communications d'entreprise, Volume X n° X/2001, pages 1à X

[6] : Competitive Intelligence Skill Cartridges User Guide (2005), TEMIS S.A. 2005

[7] : Fluck, J., C .Gieger, H.Deneke, D.Hanisch, R.Wartala, A.Holler und S.GeiBler (2003). Using Text Mining Strategies for the interpretation of Expression Data. In Proceedings of the ESF Functional genomics and disease conference, Prague, Czech Republic, May 14-17, 2003.

[8] : L.Grivel, S.Guillemin-Lanne, P.Coupet, C.Huot (2001). Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance. In Procceedings of VSST 2001, Barcelone, 15-19 octobre, 2001, FPC/UPC-SFBA-IRIT.