Système d'extraction d'information dédié à la veille
Qui est qui? Qui fait quoi? Où? Quand? Comment?

Asma BOUHAFS

Équipe Langages, Logiques, Informatique, Cognition et Communication
(LaLICC) (UMR 8139) CNRS - Université de Paris Sorbonne
96, Boulevard Raspail 75006 PARIS – France
asma_bouhafs@yahoo.com

Mots-clefs – Keywords

Classes sémantiques, Extraction d'information, Exploration Contextuelle, Ressources, Réseau sémantique.

Semantic classes, Information Extraction, Contextual Exploration, Resources, Semantic network.

Résumé – Abstract

Dans cet article nous présentons un outil d'extraction d'information dédié à la veille qui répond à un certain nombre de requêtes formulées par l'utilisateur, en combinant la puissance des outils et les ressources informatiques à une analyse linguistique. Cette analyse linguistique permet le repérage des entités nommées (acteurs, lieux, temps,…) ainsi que la mise en relation des acteurs avec leur environnement dans l'espace et le temps au moyen d'indices déclencheurs, d'indices complémentaires et de règles qui les combinent, c'est le principe de l'Exploration Contextuelle. Les résultats capitalisés dans des fichiers XML, sont proposés par le biais d'une interface, soit sous forme de graphes soit sous forme de base d'informations.

In this article we present an information extraction tool which answers a certain number of requests formulated by the user, by combining data-processing with a linguistic analysis. This linguistic analysis allows the location of the named entities (actors, places, time...) thus the relations between actors and their environments in space and time by means of indices, indicators and rules which combine them, it is the principle of Contextual Exploration. The results capitalized in XML files are presented in an interface, either in the form of graphs or in the form of databases.

1 Introduction et problématique générale

Que cela soit au niveau d'un individu, d'une entreprise ou d'une nation, anticiper les évolutions de son environnement est vital pour maintenir ou développer sa compétitivité. L'information est au cœur d'une telle démarche d'intelligence stratégique, économique et sociale. En effet, l'augmentation de la quantité d'information disponible sous forme de documents électroniques écrits en langue naturelle rend pressant le besoin de processus intelligents pour traiter de tels textes. Les méthodes de compréhension de textes ainsi que l'Extraction d'Information (EI) apparaissent comme particulièrement attrayantes et utiles. L'EI a été définie restrictivement par le programme DARPA MUC (Message Understanding Conference, 92-98), comme la tâche consistant à extraire des informations spécifiques et bien définies à partir de textes écrits en langue naturelle dans des domaines restreints, avec l'objectif spécifique de remplir automatiquement des formulaires prédéfinis ou des bases de données. Dans cet article nous décrivons un outil d'extraction d'information pour un objectif de veille: Pour mettre en évidence ces informations pertinentes à extraire, les veilleurs disposent d'outils de collecte et de traitement de l'information qui utilisent des approches traditionnelles (essentiellement statistiques et quantitatives) se révélant souvent insuffisantes. Pour être efficaces, ces outils doivent intégrer diverses techniques provenant de l'analyse linguistique, de l'intelligence artificielle et de l'ingénierie des connaissances. Ces techniques offrent un cadre méthodologique de gestion de l'information efficace et adaptable à une tâche spécifique, complexe et problématique comme la veille stratégique. Par exemple dans le domaine du terrorisme le système tentera d'extraire les noms des terroristes, des victimes, le nombre de victimes, le type d'arme utilisée, la date de l'action terroriste. Dans une société ou compagnie, elle permettrait, par exemple, de surveiller les agissements des différents protagonistes d'une transaction commerciale ou encore d'extraire les noms de compagnies impliquées, la part de capital investie par ces compagnies, le nom de la nouvelle compagnie ainsi formée et son secteur d'activité.

2 Système d'extraction d'information dédié à la veille: JavaVeille

Notre système (JavaVeille) est un système d'extraction d'information dédié à la veille, qui doit offrir un certain nombre de fonctionnalités selon le contexte d'utilisation. Il répond à un certain nombre de question du style: qui est qui? (qui sont les acteurs ou actants du domaine économique, social et politique …), qui fait quoi? (quelles sont les actions de ces acteurs, sur quels objets portent-elles et comment s'effectuent-elles ?), qui est en relation avec qui : quelles sont leurs relations avec l'environnement qui les entoure (les autres actants, leur travail…), où ? (où ont lieu les actions et les relations en question ?), quand? (quand ont-elles eu lieu ?), pour quel montant? (dans le cas du domaine économique.), etc.

2.1 Système intégrant une analyse linguistique

Lors de la conception des systèmes d'extraction d'information actuels, il est de plus en plus difficile de se passer d'une approche sémantique pour la description des significations et leur organisation théorique. La maîtrise de la gestion sémantique des informations est primordiale pour l'accès au sens d'un document. Cela impose de s'appuyer sur une organisation sémantique théorique et opératoire accompagnée d'une stratégie de recherche d'information et

d'extraction d'information porteuse de sens. Nous évoquons à titre d'exemple le réseau des catégories et sous catégories sémantiques associées aux notions "interrelation entre actants (individus, collectifs)" et "états affectant ou approuvés par un actant" (figure 1).

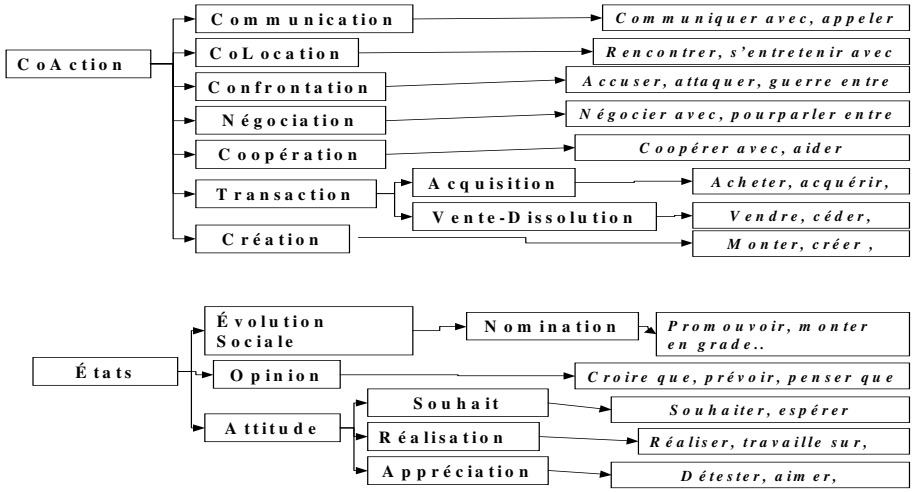

Figure 1: Réseaux d'indices des états et des relations associées aux Actants

2.2 Corpus de travail

Le repérage d'informations pertinentes nécessite de cibler précisément les informations utiles et pertinentes dans un cadre de veille. Nous avons choisi de travailler sur des textes qui développent explicitement les types de relations entre actants et leurs environnements. Un grand nombre de ces textes est disponible sur le Web. Ces dépêches articles et communiqués de presse (Le Monde, Libération, AFP…) forment des textes réels, qui ne sont pas rédigés dans le but d'y faire de l'extraction d'information. Ce sont des courtes séquences textuelles homogènes sur le plan thématique (politique, économique,…). Ce premier corpus de travail représentatif du domaine (plus de 200 articles et dépêches) contient des informations pertinentes qui peuvent être extraites et utilisées à des fins de développement.

2.3 Acquisition initiale des données linguistiques

Le travail à partir de corpus est une approche qui s'est répandue afin de permettre la mise en place de méthodes d'analyses en adéquation avec le contenu textuel source. En effet le corpus de textes ainsi constitué est une source riche en unités linguistiques qui permettent d'exprimer les relations recherchées. L'acquisition de ces données linguistiques nécessite une fouille systématique des textes en vue d'accumuler les indicateurs, les indices et les règles qui les combinent. Nous sommes partis d'un exemple (noyau) très simple exprimant la relation entre actants à partir duquel nous avons étendu progressivement l'ensemble des marqueurs qui appartiennent pratiquement à la même séquence et qui réfèrent au même type de relation ou à un type sémantiquement apparenté. Les marqueurs recensés représentent des indicateurs qui

doivent être, dans une étape suivante classés dans des classes linguistiques suivant leurs formes et leurs dispositions par rapport aux indices qui les entourent, c'est l'Exploration Contextuelle.

2.4 L'Exploration Contextuelle

L'Exploration Contextuelle (EC) (Desclés, 1997) est une méthode issue des recherches effectuées par l'équipe LaLICC[1] pour le traitement automatique des textes en langue naturelle. De nombreuses applications informatiques ont déjà été réalisées, notamment le résumé automatique (Minel, 2000), le filtrage d'informations selon différents points de vue, etc. C'est une méthode ayant une portée sémantique qui ne se base pas sur une représentation profonde du texte mais sur une identification automatique de certaines unités linguistiques (marqueurs) pertinentes pour une tâche donnée, appelées indices déclencheurs (indicateurs) et indices complémentaires. Les indices déclencheurs sont retenus en fonction d'objectifs précis (par exemple, déterminer une relation sémantique entre concepts et/ou la valeur sémantique contextualisée d'un marqueur grammatical ou lexical polysémique). Une analyse exploratoire du contexte permet d'identifier d'autres indices linguistiques, eux aussi jugés pertinents pour la tâche traitée (indices complémentaires). L'indice déclencheur et les indices complémentaires étant identifiés, ils permettent, au moyen de règles heuristiques, de prendre les décisions impliquées par l'objectif attendu dans un contexte bien défini. Ces règles se déclenchent pour attribuer à une unité lexicale (une phrase, un paragraphe, etc.) des étiquettes sémantiques, etc. Dans ce travail, nous avons appliqué l'EC dans le but d'identifier dans les textes les relations recherchées. Pour cela, nous avons construit une base d'indices linguistiques (plus de 2000 marqueurs différents) exprimant les relations entre actants et leurs environnements dans l'espace et le temps. Ces indices sont regroupés dans des classes sémantiques: celles des indicateurs et celles des indices complémentaires qui seront mises en relation par un ensemble de règles d'EC (plus de 180 règles à ce jour). L'action de l'ensemble des règles permet de construire progressivement des représentations sémantiques. Certaines règles permettent de créer des marqueurs d'EC complexes, d'autres d'attribuer une étiquette sémantique à une phrase.

3 Système informatique

Le système se décompose en quatre tâches:

1- La récupération et le traitement de l'information recueillie, (Laublet 2002), ainsi que le typage des documents par rapport à leur source (information officielle ou non,..) et à leur statut temporel (réalisé, non réalisé, en cours…).

2- Le repérage des entités nommées (personnes, compagnies, organisations, lieux, temps…) évoquées dans les documents. La reconnaissance des entités nommées est un problème qui se pose dans les différents domaines du traitement automatique de la langue naturelle (TALN) : veille technologique, indexation de textes ou traduction, extraction d'information (EI)

[1] L'équipe LaLICC : LAngages Logiques Informatique Cognition et Communication

(Poibeau T., 1999), etc. Dans notre travail la reconnaissance des entités nommées utilise une méthode qui repose à la fois sur la structure interne de l'entité nommée ainsi que sur l'analyse du contexte. Les règles utilisées sont une combinaison d'expressions régulières et d'indice lexicaux (sous la forme de règles d'EC) qui sont indépendantes du domaine. Elles permettent aussi dans une seconde étape d'identifier les expressions spatiales ainsi que les expressions temporelles qui vont être utilisées dans le traitement final.

3- Le repérage des segments porteurs d'information par l'identification de formes langagières exprimant des thématiques (certaines unités linguistiques interprétées dans leurs contextes) et des relations sémantiques entre ces entités nommées, comme les relations entre Actants (`<CoLocation>`, `<CoAction>`…) ainsi que les relations d'un actant avec son environnement (par exemple `<Localisation_spatiale>`, `<Evolution social>`, `<Opinion>`, `<Attitude>`…). En effet ce deuxième module d'annotation prend en entrée un document déjà annoté d'informations sur les acteurs d'événements, de relations spatiales et temporelles. Cette étape va utiliser des ressources linguistiques plus importantes que celles de la précédente étape. En effet certaines règles de repérage de relations entre actants s'appuient sur des segments textuels déjà annotés (`<actant>`, `<expression_temporelle>`…). Les règles d'annotation augmentent celles de la précédente étape de deux nouvelles formes : l'une qui prend en compte, dans ses prémisses, des segments textuels déjà annotés et l'autre prenant en charge les notions d'indicateur, d'indices complémentaires et d'espace de recherche selon la méthode de filtrage par l'EC.

4- La réorganisation ou la reformulation des segments porteurs d'informations pertinentes (segments étiquetés) dans le but d'obtenir un ensemble de représentations du contenu sémantique de ces segments et de traiter les informations recueillies (les sorties XML) pour pouvoir générer les représentations graphiques des résultats (Holt R. C 00) et construire et mettre à jour la base des résultats.

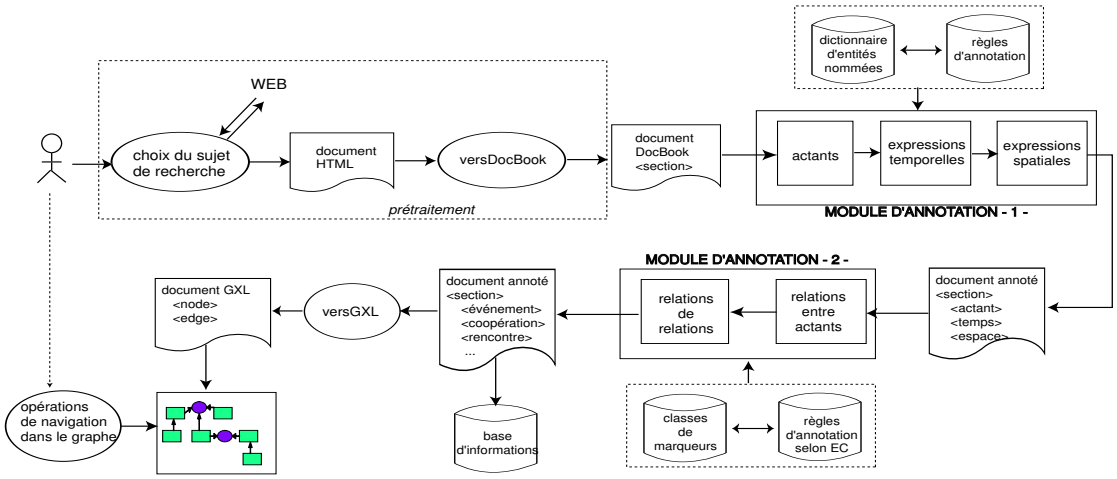

Figure 2: Architecture générale du système JavaVeille

4 Conclusion et évaluation

L'intérêt de ce travail réside dans notre volonté de fournir des outils informatiques capables d'aider à la recherche de certains types d'informations spécifiques, dans un but de veille. Le

système est évalué sur des textes qui n'ont pas servi dans l'étape d'acquisition initiale des données linguistiques. Ces textes sont ensuite incorporés dans le corpus d'acquisition et les résultats servent à enrichire notre base linguistique. Ce travail est en cours, certaines sous-tâches sont encore en cours de construction et ne sont pas encore complètement opérationnelles. Notons qu'au fur et à mesure que l'on avance dans l'implémentation, les bases de données linguistiques sont élargies de manière à pouvoir couvrir le plus grand nombre de domaines. Précisons que l'identification et la classification sémantique des indicateurs linguistiques pertinents restent une tâche importante et délicate, et que la mise en œuvre informatique exige une modélisation, où linguistes et informaticiens doivent coopérer si l'on veut aboutir à des systèmes robustes et performants. Un protocole d'évaluation est actuellement expérimenté pour permettre à la fois de cerner les limites de nos règles et d'améliorer la qualité d'extraction de notre système. Enfin soulignons que ce travail est mené dans le cadre d'une thèse de doctorat[2], et qu'il est intégré dans le cadre du projet OLETT[3].

Références

DESCLES J-P. (1997), Systèmes d'Exploration Contextuelle, dans Co-texte et calcul du sens, Claude Guimier (éd).

DOU H., JAKOBIAK F. (1995), De l'information documentaire à la veille technologique pour l'entreprise : enjeux, aspects généraux et définitions, in *Veille technologique et compétitivité*, Dunod, 1995.

GOUJON B.(2000), *Utilisation de l'exploration contextuelle pour l'aide à la veille technologique : Réalisation du système informatique VIGITEXT*. Thèse de doctorat (traitement automatique du langage naturel) : Université Paris-Sorbonne, avril 2000. 550 p

HOLT R. C., WINTER A., SCHURR A. (2000). Gxl: Towards a standard exchange format. In *Proceedings 7th Working Conference on Reverse Engineering (WCRE 2000).*

LAUBLET P, NAIT-BAHA L, JACKIEWICZ A, DJIOUA B, (2002) *"Collecte d'information textuelles sur le Web selon différents points de vue* ", Céline Paganelli (ed), Interaction homme-machine et recherche d'information (chapitre 8), Hermes , , 2002.

MINEL J-L., DESCLES J-P., CARTIER E., CCRISPINO G., BENHAZZEZ S., JACKIEWICZ A., (2000), *"Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText* ", revue TSI.

POIBEAU T., (1999) Repérage des entités nommées : un enjeu pour les systèmes de veille. In *Actes des troisièmes rencontres de Terminologie et Intelligence Artificielle (TIA'99)*, volume19, p. 43–51.

---

2 Thèse de doctorat dirigée par M. Desclés à Paris IV

3 OLETT est un projet soutenu par le programme interdisciplinaire du CNRS société de l'information et porte sur l'identification des événements et des lieux pour l'organisation aspecto-temporelles sous-jacente au texte (application : filtrage sémantique automatique, recherche et extraction d'informations sur le Web). Il menée par les équipes LaLICC de Paris 4 La Sorbonne et Lattice Paris 7, et financé par les départements STIC (Sciences et Technologies de l'Information et de la Communication) et SHS (Sciences de l'homme et de la société) du CNRS.