

White Paper Lingway

Pas de structuration de textes sans traitement automatique des langues

**Extraction d'informations,
Génération de méta-données,
et Indexation automatique**

Janvier 2005

>lingway

Solutions in language processing

Sommaire

1	L'EXTRACTION D'INFORMATIONS : UN BESOIN NOUVEAU.	3
2	LES METHODES D'EXTRACTION D'INFORMATION.....	4
2.1	LA PHASE DE COMPREHENSION : L'ANALYSE LINGUISTIQUE.....	4
2.2	LA PHASE DE SELECTION	5
3	L'EXTRACTION D'INFORMATIONS DANS LINGWAY KM	6
3.1	LE REPERAGE DE LA STRUCTURE LOGIQUE DU DOCUMENT	6
3.2	EXTRACTION D'ENTITES NOMMEES	7
3.3	EXTRACTION DE DESCRIPTEURS THEMATIQUES.....	7
3.4	L'EXTRACTION DE PHRASES IMPORTANTES	7
3.5	L'EXTRACTION D'ATTRIBUTS ET DE FAITS.....	8
3.6	EXTRACTION DE TERMINOLOGIES MULTILINGUES.....	8
3.7	EXTRACTION D'ASSOCIATIONS ENTRE OBJETS EXTRAITS.....	9
4	CONCLUSION	10

1 L'extraction d'informations : un besoin nouveau

Le besoin de structurer l'information textuelle est aussi ancien que la problématique documentaire : méthodes d'indexation, langages documentaires, thésaurus, ont pendant longtemps été les outils de base des documentalistes. A ces approches d'indexation manuelle des documents, lentes et coûteuses et donc applicables seulement à des bases de données de taille raisonnables, on opposait souvent les systèmes de recherche «full text» qui faisaient l'économie d'une indexation thématique et allaient rechercher directement dans les textes, ce qui permettait de traiter des volumes importants, quitte à perdre en qualité de recherche.

L'évolution majeure de ces dernières années est que l'étape de l'indexation proprement dite est de plus en plus automatisable. Cette évolution se fait sous la pression de plusieurs facteurs :

- Le volume sans cesse croissant de données à traiter oblige à chercher des méthodes automatiques.
- L'évolution liée au Web Sémantique, qui tend à structurer de plus l'information, avec l'apparition de nouveaux langages (RDF, OWL), et d'ontologies qui visent à permettre l'interopérabilité entre sources documentaires.¹
- La demande croissante d'outils liés à l'intelligence économique, qui nécessite des outils plus précis que les moteurs booléens ou statistiques classiques, pour mieux repérer les informations intéressantes, et pour repérer des associations nouvelles entre informations.
- Le besoin croissant de traiter une information multilingue, qui nécessite des outils d'extraction de correspondances entre langues.
- La demande croissante d'outils d'aide à la lecture, conséquence du volume croissant de documentation textuelle disponible : navigation rapide dans un texte, résumé automatique, identification des passages importants sont autant de fonctionnalités qui ne peuvent fonctionner correctement que sur des textes analysés.

Les techniques d'extraction d'information (« **IE pour Information Extraction** »), basées sur des technologies du traitement automatique des langues (TAL) permettent de structurer une information textuelle qui est au départ dépourvue de toute structure logique. Le champ de l'IE est souvent décomposé en plusieurs sous problèmes qui sont :

- l'extraction d'entités nommées
- l'extraction de descripteurs thématiques (libres ou normalisés)
- l'extraction de phrases importantes sous un point de vue donné
- l'extraction d'attributs
- l'extraction d'associations entre entités nommées et descripteurs
- l'extraction de correspondances multilingues

Des problématiques plus «pointues», comme la résolution de coréférences sont maintenant souvent rattachées à l'IE. Bien que ce dernier type de problème n'ait pas encore de solution générale, Lingway participe à des travaux de recherche sur ces sujets.

¹ Voir sur ce point « *Pas de Web sémantique sans traitement automatique des langues* », white paper publié sur www.lingway.com

Enfin, bien qu'il ne s'agisse pas à proprement parler d'extraction d'informations, les méthodes de structuration logique du texte – repérer le titre, les différentes parties, etc. – sont assez voisines et font appel aux mêmes types de techniques.

2 Les méthodes d'extraction d'information

Extraire de l'information d'un texte suppose deux types de traitement : une phase de « **compréhension** » et une phase de « **sélection** ». Certains outils purement statistiques font l'impasse sur la phase de compréhension, ce qui conduit généralement à des résultats discutables.

En effet, les méthodes traditionnelles qui ne reposent pas sur des techniques linguistiques ont tendance à trouver « trop » de descripteurs, notamment parce que des descripteurs ayant le même sens, qui ne sont que des variantes l'un de l'autre, ne sont pas unifiés.

Dans Lingway, l'extracteur de *thèmes ou descripteurs (objets documentaires)* se base sur la reconnaissance de *termes (objets linguistiques)* qui sont repérés par leur structure linguistique propre : par exemple en français un terme formé sur la structure « Nom + préposition + Nom + Adjectif ? » comme « *augmentation des salaires* » ou « *augmentation du salaire minimum* » est un bon candidat de descripteur thématique. Il sera effectivement retenu ou non pour indexer le document sur la base de critères complémentaires.

Si l'extracteur trouve deux descripteurs comme « *action du gouvernement* » et « *action gouvernementale* », ils seront automatiquement unifiés sous une des deux formes, ce qui constitue évidemment une indexation beaucoup plus efficace que si les deux formes étaient retenues en parallèle.

Cette unification peut être faite sur la base de critères morpho-syntaxiques, comme dans l'exemple ci-dessus (« *gouvernemental* » est un adjectif qui dérive du nom « *gouvernement* »), mais peut également être faite sur des critères sémantiques, à partir des données du dictionnaire (« *action de Matignon* ») pourrait ainsi être unifié avec « *action du gouvernement* ».

2.1 La phase de compréhension : l'analyse linguistique

Des progrès ont pu être faits dans le domaine du TAL à partir du moment où l'on a pu décomposer le problème, trop vaste pour être abordé globalement, en plusieurs sous problèmes, correspondant à plusieurs sous tâches articulées entre elles, cette décomposition étant faite à la fois au niveau des ressources linguistiques (dictionnaires et règles de grammaire) et au niveau des algorithmes (analyseurs, transducteurs, etc.).

Les principaux niveaux classiques sont :

- le niveau morphologique : identification des mots d'une phrase,
- le niveau syntaxique : identification des constituants et des fonctions d'une phrase,
- le niveau sémantique : identification du sens des mots et de la structure logique d'une phrase,

- le niveau du texte : identification des relations entre les phrases et de la structure d'un texte,
- le niveau du corpus : identification de la structure d'un ensemble de textes.

Chaque tâche (niveau) peut à son tour être décomposée en sous-tâches.

Ainsi le niveau morphologique peut se re-décomposer en :

- la tokenisation : identification des frontières de mots (simples et composés), et de phrases,
- le tagging : identification de la catégorie - nom, verbe, adjectif - de chaque mot,
- la lemmatisation : identification de la forme canonique dans le dictionnaire.

Le niveau syntaxique est souvent décomposé en :

- le chunking : identification des frontières majeures de constituants (groupe nominal, verbal, etc.) et/ou des relations majeures entre les mots,
- le tagging fonctionnel : affectation de fonctions grammaticales aux constituants,
- le parsing : construction d'un arbre représentant la structure de la phrase complète.

Le niveau sémantique se décompose également en :

- la sélection de sens (WSD « word sense disambiguation ») : choix du sens de chaque mot,
- la structuration logique : identification des arguments de chaque prédicat et de leur rôle sémantique (agent, but, lieu, etc.).

Le niveau texte peut aussi se décomposer en :

- la résolution des anaphores (antécédents des pronoms, ellipses, références),
- la détermination de la structure rhétorique (commentaires, explications, causalités, etc.),
- la détermination de la structure thématique (de quels sujets le texte traite-t-il ?).

Enfin le niveau du corpus peut se décomposer en :

- la détermination de la nature des documents (article de presse, article technique, texte réglementaire, brochure commerciale, etc.),
- la structure thématique du corpus (de quels sujets le corpus traite-t-il ?).

La plupart des ces tâches correspondent à des sous disciplines de la linguistique. Du point de vue de l'ingénierie du TAL, elles correspondent à des modules, des programmes spécifiques qui ont chacun besoin d'un type de ressources linguistiques et d'un type d'algorithmique particulier.

Tous ces modules sont, à des niveaux divers, implémentés et utilisés dans les outils Lingway.

2.2 La phase de sélection

Dans de nombreux cas d'application, la phase d'extraction ne suffit pas. Il faut encore trier entre ce qui est intéressant et ce qui ne l'est pas, ce qui dépend généralement du contexte d'utilisation.

Dans cette phase, on se sert en premier lieu d'indicateurs statistiques sur la fréquence des mots dans la langue générale, afin de donner plus ou moins d'importance à des termes extraits et ainsi ne garder que les plus pertinents. Par exemple, dans le paragraphe qui précède, le terme « cas d'application » (Nom+Préposition+Nom), correspond au motif indiqué au début de cette section. Il sera donc candidat pour être descripteur du texte. Mais comme le nom «cas» est très

fréquent en langue française, le terme aura un « score » faible et est candidat à être rejeté (en première approche) par la phase de sélection au profit de termes plus porteurs d'information.

Il s'agit toutefois d'un premier niveau de sélection qui doit être affiné en fonction du texte, voire du corpus. En effet, dans un texte de spécifications informatiques où le terme « cas d'application » serait utilisé comme traduction de « use case », ce terme serait effectivement pertinent. Ici, d'autres indicateurs statistiques basés sur la fréquence dans le texte lui-même ou le nombre de co-occurrences des deux mots (« cas » et « utilisation ») vont venir modifier le score et conduire à considérer « cas d'application » comme un bon descripteur à retenir pour ce document.

Enfin, le corpus lui-même joue un rôle de modulation à un niveau supérieur, en utilisant des techniques similaires. Pour reprendre notre exemple, si le corpus étudié est une compilation de tous les « use case » d'une société, alors le terme devient finalement inintéressant et sera écarté. S'il s'agit d'un ensemble plus vaste, alors le terme permettra de rapidement sélectionner un sous-ensemble de document.

D'une manière générale, moins de 10% des termes extraits sur des bases linguistiques sont retenus comme descripteurs pertinents par Lingway KM. Cette phase de sélection est donc très importante dans le processus. Mais il est crucial, pour que de tels indicateurs statistiques « fonctionnent », qu'ils s'appliquent à d'un traitement linguistique avancé : d'une part, une sélection ne peut créer de « bons » termes si elle s'applique sur une mauvaise extraction ; d'autre part les mesures statistiques seront très différentes suivant que l'on considère, par exemple, « action gouvernementale » et « action du gouvernement » comme des entités différentes sans rapport l'une avec l'autre ou non. Là où des comptages sur des chaînes de caractères se révèlent de piètre qualité dans nombre de systèmes non-linguistiques, ils donnent toute leur puissance sur des entités linguistiquement et sémantiquement fondées.

C'est le couplage d'algorithmes statistiques avancés à une extraction linguistique puissante qui fait la force de Lingway KM.

3 L'extraction d'informations dans LINGWAY KM

Les paragraphes suivants décrivent les types d'informations extractibles par Lingway KM, et la façon dont elles sont utilisées.

3.1 Le repérage de la structure logique du document

En utilisant à la fois la structure du texte (balises HTML, typographie, etc.) et des règles d'analyse du contenu textuel, le système peut repérer :

- des séquences comme un résumé, une introduction, une conclusion, une bibliographie, des remerciements, une annexe ; etc.
- la présentation d'un résultat, une comparaison avantages / inconvénients, une comparaison de prix, etc.

Ces modules sont généralement spécifiques d'un type de document donné (un article scientifique, une circulaire administrative, une offre d'emploi, etc.).

3.2 Extraction d'entités nommées

L'identification d' «entités nommées» a été identifiée comme une tâche à part entière lors des campagnes d'évaluation MUC (Machine Understanding Conference) il y a une dizaine d'années. On distingue généralement trois sous tâches : reconnaissance des noms d'entités (organisations, personnes, lieux) , des expressions temporelles (dates et autres désignations temporelles) et des expressions numériques (grandeurs mesurables, quantités, pourcentages, etc.).

Lingway permet d'extraire automatiquement ce type d'entités

- personnes: essentiellement sur la base de règles de type « prénom + X » ou « X président de Y ». L'extracteur de nom de personnes identifie le nom patronymique, le prénom, une fonction quand elle est présente.
- lieux : basé essentiellement sur des fichiers de noms de lieux. L'extracteur identifie le nom du lieu, son type (pays, ville, région, etc.).
- organisations : basé sur des listes et des règles. L'extracteur identifie le nom de l'organisation et son type (société anonyme, association, etc.)
- dates et autres marqueurs temporels.

3.3 Extraction de descripteurs thématiques

On distingue habituellement plusieurs types d'indexation thématique :

- Indexation contrôlée : seuls des termes appartenant à une liste normalisée dite « d'autorité » pré-établie peuvent être retenus. On utilise souvent aussi le vocable de « catégorisation » pour désigner ce type de traitement.
- Indexation libre : les termes choisis comme descripteurs thématiques peuvent être librement choisis, sans contrainte par rapport à une liste d'autorité.

Les problèmes qui se posent quand on veut automatiser l'indexation d'un texte en descripteurs thématiques sont sensiblement différents selon que l'on utilise l'un ou l'autre de ces modes d'indexation. Mais dans les deux cas on a d'abord une phase d'identification des « thèmes », puis un éventuel traitement de regroupement ou de normalisation de ces termes, et enfin un calcul de pertinence indiquant l'importance du thème dans le document.

Cette normalisation peut se faire dans le cadre d'une indexation libre ou contrôlée.

3.4 L'extraction de phrases importantes

Les phrases importantes sont repérées par la présence de « marqueurs linguistiques » typiques – c'est à dire des séquences de mots typiques d'un type de phrase –, éventuellement complétée par l'identification de la partie de texte dans laquelle se trouve la phrase, et enfin de mots importants dans le contexte. (par exemple, on peut tenir compte du fait qu'un mot du titre est cité dans la phrase considérée).

Des exemples de phrases importantes sont par exemple

- Une annonce thématique, c'est à dire une phrase annonçant ou reprenant les ou les thèmes traités dans le document ou dans une partie donnée. Par exemple : « **Dans ce texte, nous étudierons les implications de nos travaux et nanotechnologies** », « **Notre document se divise en 3 parties, ...** »
- Un renforcement, c'est à dire une phrase dont l'auteur, par des moyens rhétoriques, cherche à souligner l'importance. Par exemple : « **Les fullerenes ont une structure particulièrement importante pour les applications médicales** », « **J'insiste sur le fait que ces techniques n'ont pas d'effets secondaires...** »

Un important travail de phraséologie a été entrepris pour décrire ces marqueurs linguistiques, généralement par rapport à un type de corpus donné. Le point important est que même s'il peut exister pour un type de phrase donné un grand nombre de marqueurs, ils sont dans la plupart des cas en nombre finis et donc descriptibles.

Une méthodologie a été mise au point qui permet, en utilisant divers outils d'analyse du corpus et les dictionnaires généraux, d'aider à la construction de ces listes.

3.5 L'extraction d'attributs et de faits

L'extraction d'entités permet d'identifier des personnes, des organisations, des lieux, etc... Mais on veut généralement aller plus loin et associer à ces entités des attributs extraits du texte : la *fonction* d'une personne (directeur, chef de projet, etc.), son *rôle* dans un document (auteur, personne citée, personne en charge d'un dossier, destinataire, etc.). Ou encore la *nature* d'une organisation (société, association, groupement, etc.), son *chiffre d'affaires*, etc.

Dans Lingway, des règles contextuelles permettent d'extraire ces associations et de produire une description XML/RDF appropriée.

Par exemple, pour une application dans le domaine de l'emploi, Lingway permet d'identifier dans un texte d'offre d'emploi divers décrivant l'offre comme la *taille* de l'entreprise, le *motif* de l'embauche, l'*expérience* demandée, le *rattachement hiérarchique* proposé, etc.

3.6 Extraction de terminologies multilingues

Lingway KM permet de construire automatiquement une terminologie bilingue, sans nécessiter un corpus aligné, comme c'est généralement le cas dans d'autres approches. Cette fonctionnalité est particulièrement utile lorsque Lingway KM est utilisé conjointement avec un moteur de traduction automatique.

Le principe général est le suivant :

Les termes sont extraits dans le corpus en langue source. Chaque terme est envoyé comme une question « cross-langage » à un corpus langue cible, qui aura été préalablement constitué. Ce corpus doit évidemment être dans le même domaine que le corpus en langue source, mais n'a pas besoin d'être aligné.

Les termes correspondants dans le corpus langue cible sont autant de candidats à être des traductions. Celui –ou ceux- qui apparaissent le plus fréquemment sont retenus comme des traductions des termes de départ.

Cette méthode permet de construire très rapidement des terminologies de plusieurs milliers de termes, ce qui est particulièrement intéressant dans divers cas de figure :

- Génération de dictionnaire pour un moteur de traduction.
- Catégorisation multilingue, ouvrant la possibilité de catégoriser dans d'autres langues que celle utilisée lors de l'apprentissage.
- Construction de vocabulaires d'entreprise

3.7 Extraction d'associations entre objets extraits

La demande pour des outils de cartographie documentaire est de plus en plus marquée. Une des fonctionnalités généralement demandée est de repérer des associations entre des entités nommées et/ou des descripteurs thématiques identifiés dans les textes. Pour que ce type d'approche soit utile, encore faut-il que les objets extraits soient signifiants : on voit certains outils de cartographie qui font des rapprochements entre des mots simples, ce qui n'a que peu d'intérêt.

Le graphe ci dessous montre des associations extraites dans la presse fin 2004 : on note comme par exemple les descripteurs suivants qui sont particulièrement intéressants :

- mort de Yasser Arafat
- mort officielle de Yasser Arafat

On se souvient en effet que le décès de Yasser Arafat avait été largement évoqué avant d'être officiellement annoncé.

On voit également les descripteurs

- relance du processus de paix
- nouvelle direction palestinienne,

qui sont liés à « mort de Yasser Arafat ».

Ce type d'association ne peut être extrait que si les termes sont « sémantiquement pleins », et donc le résultat d'une analyse linguistique. Sinon, on risque de faire apparaître un lien entre « mort », « Yasser » et « relance », ce qui n'est pas particulièrement utile.

