Bénéfice d'un catalogue spécialisé de sites internet médicaux pour la constitution de corpus à des fin de recherche.

Thierry Delbecque, Pierre Zweigenbaum

20 janvier 2006

1 Contrôler les sources de documents web en s'appuyant sur des catalogues existants

Internet aujourd'hui constitue une vaste ressource documentaire, qu'il est possible d'envisager à des fins diverses intéressant le TALN (construction de corpus, extraction de ressources terminologiques, etc.). Cependant, l'un des cotés précieux d'internet, à savoir la liberté pour chacun de publier à volonté sur la toile, pose différents problèmes.

Par exemple, pour la constitution de corpus spécialisés, il peut être souhaitable de savoir collecter un ensemble de documents en maîtrisant en même temps la couverture thématique de l'ensemble finalement construit, et le niveau de langage utilisé. En fonction des caractéristiques de chaque document qui les composent, les corpus peuvent conduire les outils de TAL à produire des résultats variés ; afin de contrôler ou de comprendre cette variabilité il faut que la construction du corpus prenne en compte les caractéristiques de chaque source qu'il inclue [1]. D'autre part, l'utilisation de moteurs de recherches généralistes tels que Google, dont les critères de tri dépendent du "taux de popularité" des sources indexées, peut conduire à la construction de ressources biaisées ou incomplètes. Enfin, concernant un aspect purement "extraction d'information", la validité de l'information contenue dans les documents devrait pouvoir être certifiée, en particulier pour des systèmes de Question/Réponses.

Chacun de ces éléments souligne l'intérêt tout particulier des portails documentaires spécialisés, maintenus par des professionnels du domaine. Pour la médecine, le portail CISMeF (Catalogue et Index des Sites Médicaux Francophones, [2]) est un exemple de tels sites, qui permet d'apporter des réponses aux points soulevés plus haut. En particulier, CISMeF a été utilisé dans le cadre de la compétition EQueR (évaluation de systèmes de Questions Réponses, [3]) afin de constituer la base documentaire de laquelle les systèmes concurrents devaient extraire les réponses aux questions suggérées par les

organisateurs de l'évaluation. Nous avons participé à EQueR à la fois pour la mise en place du corpus utilisable par chacun des concurrents, puis pour certains auteurs en tant que participants [4]. Dans ce contexte, nous avons été amenés à effectuer différentes expériences d'indexation sur ce corpus, et différentes mesures statistiques, qui nous ont permis de faire apparaître une cartographie thématique des documents indexés, et d'envisager l'influence de cette cartographie dans des systèmes d'extraction d'information.

La présentation résumée ici entend relater cette expérience ainsi que ses résultats. Comme sujet de discussion, nous présenterons les travaux en cours actuellement sur ce corpus, ainsi quelques perspectives que ces travaux nous inspirent, en ce qui concerne l'aspect sous lequel les documents sont accessibles, en regard des difficultés que nous avons rencontrées. Cela éclairera encore davantage l'importance que revêtent les catalogues professionnels de sites spécialisés sur la toile pour la constitution de corpus.

2 L'expérience

2.1 Définition et extraction du corpus

Le corpus est un sous-ensemble des documents indexés par CISMeF, faisant partie d'une dizaine de sources spécialisées différentes, et auxquels viennent se joindre les documents référencés *un lien plus loin*. L'ensemble comporte 5147 documents, et la présentation orale en précise la source exacte, ainsi que les traitements que nous avons dû effectuer afin d'obtenir une ressource exploitable pour la compétition et pour des tâches automatiques de TALN [5]. Le corpus obtenu dispose de plusieurs atouts, dont une cartographie thématique donnée a priori du fait de la vocation claire de chacun des sites sélectionnes (public visé, vocation plutôt médicale versus plutôt réglementaire, etc.).

2.2 Les expériences sur le corpus : indexation sémantique et extraction d'information

Nous avons effectué une expérience d'indexation du corpus, afin de voir s'il était possible d'obtenir une cartographie thématique comparable à la cartographie donnée a priori par la nature des domaines sources des documents. Une voie possible aurait été la voie lexicométrique basée sur la représentation vectorielle de chaque document dans l'espace des termes. Nous avons choisi de mettre en œuvre une ressource termino-ontologique du domaine, l'UMLS [7], qui est une agrégation de ressources terminologiques médicales, organisées par des ensembles de relations entre termes, et enrichie d'un réseau sémantique. Après avoir effectué le repérage des concepts de l'UMLS au sein du corpus en mettant en œuvre une plate-forme développée spécialement pour cela (phase d'indexation), des analyses factorielles nous ont permis de

faire apparaître une organisation des documents du corpus cohérente avec la vocation de chacun des documents en fonction de sa source (thématique a priori) [6].

Suite à cela, la deuxième expérience que nous avons menée a porté sur un aspect lié aux systèmes de questions réponses. En se focalisant sur un type particulier de questions médicales, et sur la base d'analyses purement statistiques, nous avons cherché à mettre en évidence l'efficacité ou non des concepts de l'UMLS projetés sur le corpus comme indice afin d'extraire les fragments de documents où pouvait figurer la réponse. L'un des résultats de cette analyse a démontré l'importance de l'origine du document dans le comportement de l'indexation UMLS vis-à-vis de cette tâche. Nous avons également fait apparaître un lien entre l'efficacité de l'indice (concept UMLS) dans la tâche d'extraction de réponse et la cartographie calculée dans l'expérience précédente [5].

3 Discussion/conclusion : les travaux actuels, les difficultés

Nous continuons de travailler sur ce corpus, toujours dans un objectif de recherche sur les systèmes de question réponses, et les systèmes d'extraction d'information en général. Nos travaux actuels mettent davantage à contribution des analyses linguistiques (analyses syntaxiques et grammaticales des textes) que lors de la compétition EQueR, avec toujours l'ambition d'utiliser des méthodes géométriques pour exploiter la structure du corpus, mais également des méthodes issues de l'apprentissage automatique.

Chaque étape des traitements que nous effectuons sur le corpus consiste à enrichir celui-ci en informations sémantiques, syntaxiques, ou structurelles. D'autres données importantes concernent la caractérisation globale du document au sein de l'ensemble total par le biais de quelques grandes dimensions telles que le type (article, document pédagogique, ...), l'auteur, le public visé, etc. [1]. En fait chaque étape ajoute au document original une nouvelle couche d'information soit orthogonale aux couches précédentes, soit qui vient les prolonger. Chacune de ces couche est un objet qui doit être parfaitement aligné avec le document initial ; de plus, pour certaines d'entre elles, leur intérêt peut être général, en ce sens que d'autre groupes de travail pourraient en bénéficier (par exemple, l'indexation par des concepts de l'UMLS, ou le résultat d'analyses syntaxiques).

Cela pose la question de la structure des documents disponibles sur internet. La plupart du temps, ceux-ci sont fournis au format HTML (au mieux), ou PDF. Outres les difficultés inhérentes à l'extraction du texte, ce dernier format est dédié à la présentation, et est dénué de toute information sémantique sur son contenu ; les balises HTML ont elles-mêmes vocation à supporter la mise en page, plutôt que la structure logique du document. Après

un travail préalable d'extraction du contenu textuel, souvent peu gratifiant et dont le résultat n'est pas toujours entièrement satisfaisant (en particulier dans le cas des documents PDF), l'une des premières tâches que nous effectuons est de transformer le texte en un document XML, défini de manière ad-hoc selon nos besoins. La suite des traitements est en fait une suite de transformations, par enrichissement et filtrages, de fichiers XML.

Cela pointe une fois de plus le bénéfice qu'il y aurait à disposer directement de documents accessibles dans un format plus formel et communément défini, par l'intermédiaire d'un dialecte XML comme par exemple celui du standard XCES de la Text Encoding Initiative [1, 8] pour l'échange de corpus, qui offrirait les informations "de base" (structuration du texte en sections, normalisation des tableaux, identification des entités élémentaires, etc.) ; ceci faciliterait l'alignement des différentes couches d'information avec le document, et ouvrirait la possibilité de mettre en commun les différentes couches d'informations générées par différentes équipes sur ces documents. Certainement inenvisageable à grande échelle, ceci devrait être possible dans le cadre de portails spécialisés, que l'on pourrait alors considérer en plus comme des coordinateurs de ces nouvelles ressources et des ressources dérivées, ce qu'ils sont vraisemblablement les seuls à pouvoir faire.

Mots clés

Internet, catalogues, extraction d'informations, XML, travail coopératif

Références

[1] Zweigenbaum P., Jacquemart P. Grabar N., Habert B. Building a text corpus for representing the variety of medical language. MEDINFO 2001.

[2] Darmoni S., Thirion B., Leroy J.P., Douyère M., Baudic F., Piot J. CISMeF : a structured Health resource guide for healthcare professionals and patients. RIAO 2000.

[3] Ayache C., Grau B., Vilnat A., Campagne d'évaluation EQueR-EVALDA : évaluation en questions-réponses. TALN 2005.

[4] Delbecque T., Zweigenbaum P., Berroyer J.F., Poibeau T. Le système STIM/LIPN à EQueR 2004, tâche médicale. TALN 2005.

[5] Delbecque T., Jacquemart P. Zweigenbaum P. Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales. CORIA 2005.

[6] Delbecque T., Zweigenbaum P. Indexation UMLS en français : une expérience. JFIM 2005.

[7] www.nlm.nih.gov/research/umls/UMLSDOC.HTML

[8] www.xml-ces.org