Saadi Lahlou and Helka Folch

Quelques stratégies pour l'exploitation de gros corpus en analyse des données textuelles

Book section

(1998) - LAHLOU, Saadi, FOLCH, Helka. Quelques stratégies pour l'exploitation de gros corpus en analyse des données textuelles. 4èmes Journées Internationales d'Analyse des Donnés Textuelles. Nice, France, 18-21 février 1998. pp. 381-390.

Quelques stratégies pour l'exploitation en ADT de grands corpus hétérogènes

Saadi LAHLOU* et Helka FOLCH**

(*EDF/DER. 92141 Clamart Cedex. *saadi.lahlou@der.edfgdf.fr*)

(**ENS Fontenay-St Cloud. 92266 Fontenay aux Roses. *folch@ens-fcl.fr*)

Our work carried out as part of the Scriptorium project has confronted us with a variety of problems that have to be faced by analysts engaging in text-mining as applied to large heterogeneous corpora (intranet, www, document-based DB). This paper presents several solutions concerning the following points : the extraction of relevant sub-sections of the corpus, meta-data, efficient storage, historisation.

We introduce two original solutions : document storage based on collections of self-describing texts with embedded meta-data in the form of mark-up (instead of a DBMS or file-based approach : full text indexing at such a scale is heavy) ; use of an extractor based on the software product TOPIC to retrieve relevant paragraphs and assemble them into homogeneous sub-corpora of exploitable size (< 10 Mega).

We shall also describe the strategies we have adopted for comparing different analyses of the corpus in a historical perspective, in particular the transformation of ALCESTE class profiles into TOPIC concepts aimed at providing fixed, quantifiable measurements of the density of certain topics in the texts.

1. Qu'est-ce que Scriptorium ?

SCRIPTORIUM [3] est un dispositif sociologique qui cherche à repérer les points de vue des acteurs d'une grande entreprise à partir des traces documentaires (plans stratégiques, comptes rendus de réunions, bilans sociaux, procès-verbaux de comités paritaires, tracts, journaux internes etc.). Les prises de position des acteurs laissent des traces discursives, qui révèlent leurs représentations. Il s'agit de constituer une base de textes à partir d'un échantillon raisonné des discours, sorte de « mémoire sociale » ; puis, à l'aide de méthodes d'ADT, d'en observer les variations dans le temps et l'espace social pour faire émerger les tendances.

Le projet, démarré en 1995, se déroule maintenant en coopération entre la Direction des Etudes et Recherches d'EDF (département GRETS) et l'ENS Fontenay-Saint Cloud (équipe ELI). Il inclut une chaîne complète de traitement : alimentation (OCR ou import, formatage), marquage, indexation et stockage des documents, extraction et traitement par des outils

logiciels, et l'analyse proprement dite. Les résultats présentés ici proviennent d'une « maquette » de test, de quelque 50 Mega-octets en format .txt [1], soit 7,5 millions de mots ou 135000 paragraphes. Celle-ci contient des textes océrisés (ouvrages, rapports, articles) centrés sur le thème du service public, mais également quelques belles séries chronologiques de discours produites par les acteurs sociaux de l'entreprise (10 ans sur une base infra-hebdomadaire), des réponses de milliers d'agents à des questions ouvertes, des P.V. de réunions paritaires etc. La composition du corpus est la suivante : articles de périodiques (12 M), retranscription de tracts ou communiqués syndicaux (10 M), livres (5 M), communiqués et dossiers de presse (2 M), lettres d'information (7 M), notes et rapports (4 M), études (3 M), enquêtes ouvertes (2 M), autres (8 M). Il s'agit donc d'une base homogène sur le plan des acteurs et thèmes, mais hétérogène sur le plan des condition de production des discours.

L'architecture organise, autour d'un système de stockage, un ensemble d'outils applicatifs en entrée (alimentation) et sortie (recherche, extraction, analyse, visualisation). Actuellement, SCRIPTORIUM comprend quatre modules fonctionnels :

Alimentation : Elle se fait par : scannage + OCR de documents papier, extraction de CD-ROM, saisie directe, téléchargement sur Intranet et Internet.

Stockage : Les documents sont stockés dans la base dans leur format d'origine, et en « TXT ». Ils sont accompagnés d'un "cartouche" renseignant l'origine et le contexte du document. Ces informations sont implémentées sous forme de balises SGML. Après conversion au format « TXT » et sgmlisation, les documents sont indexés automatiquement en texte intégral.

Recherche/Extraction : L'extraction est réalisée par un moteur de recherche plein texte à partir de combinaisons booléennes de mots clés, d'expressions ou par filtrage sur les champs du cartouche. Le grain d'extraction, dans la version actuelle, est le paragraphe. L'objectif de ce module est d'une part, de produire des « coupes » du corpus en fonction de critères multiples, d'autre part d'attribuer un score aux unités élémentaires de texte.

Analyse . Les fragments de texte produits par l'extracteur sont traités par des méthodes d'ADT visant à dégager les pôles de sens par repérage des champs lexicaux. D'autres traitements statistiques non textuels sont faits sur des matrices numériques ayant comme individus les fragments extraits et comme variables les champs signalétiques et les scores.

Pour l'objectif, qui est de tracer les trajectoires discursives et temporelles des différents acteurs dans un espace sémantique pour mieux comprendre leurs prises de positions, il est indispensable de disposer de corpus homogènes, comparables, pertinents (centrés sur les thèmes étudiés [4]), en repérant les sources par un signalétique standardisé. L'hétérogénéité et le volume du corpus qui sert de minerai de base, qui excèdent parfois, déjà sur la maquette, les capacités actuelles de nos logiciels (notamment Alceste [7]), posent des problèmes extrêmement intéressants. On en présente ici qui concernent l'extraction, la signalétique et le stockage, l'historisation, la construction de concepts ; et les stratégies adoptées en réponse. Ces problèmes seront rencontrés sous une forme ou une autre par tous les mineurs de texte.

2. Extraire des sous-corpus pertinents et homogènes

La pertinence et l'homogénéité du corpus sur lequel vont opérer les logiciels d'ADT sont des facteurs essentiels de l'interprétabilité des résultats et de la qualité des analyses. Notre base est hétérogène : elle comprend des documents de taille différente, produits dans des conditions différentes. Chaque analyse impose d'opérer sur une "coupe" du corpus, i.e. une extraction de fragments de texte homogènes suivant certains paramètres. Ces paramètres peuvent porter sur les sources et/ou les conditions d'énonciation (ex. : série chronologique de lettres ouvertes de la direction), ou le contenu (ex. : paragraphes contenant le concept de "négociation"), ou les deux (ex. : textes syndicaux, sur la notion de service public, durant telle période).

Les critères de coupe concernant le para-texte ("hors-texte" ou "données mondaines": qui a parlé, quand, où, à qui, pourquoi, comment etc.) sont relativement simples à prendre en compte. Cela revient à faire une première extraction des textes produits dans des conditions déterminées. Sachant que ces données mondaines sont décrites dans un cartouche (fiche signalétique) attaché à chaque texte, l'extraction peut se faire avec des outils de bas niveau, ou directement à partir d'un extracteur de type recherche documentaire.

La détermination de coupes à partir des données mondaines est facile ; mais elle fournit des textes de taille inégale, et surtout inégalement riches en la matière qui nous intéresse si nous visons un thème particulier. Pratiquement, les acteurs ont tendance à aborder différents thèmes lors de la production d'un texte : ils rentabilisent l'occasion de parole en traitant plusieurs questions à la fois. Si l'on se contente de travailler sur des corpus composés de textes entiers, on récupère surtout par l'analyse le registre des thèmes et des rhétoriques utilisés par les acteurs considérés ; mais le découpage fin des différents aspects d'un thème particulier est écrasé par l'hétérogénéité du corpus. Une des principales difficultés est donc la taille du grain d'extraction. Quand on s'intéresse à un thème particulier (ex. : le service public), l'analyse sur les textes entiers donne des indications sur la position de ce thème dans le discours plus général des acteurs ; par contre, pour examiner les différents aspects du thème lui-même, il faut un grain d'extraction plus fin, pour trier dans le texte tout venant le minerai intéressant (par exemple les paragraphes qui contiennent ce concept ; c'est l'unité que nous avons utilisée).

Pratiquement, nous avons utilisé le logiciel TOPIC, un outil de recherche documentaire, développé par la société VERITY, qui comprend trois utilitaires : mkvdk (utilitaire d'indexation), TDE (moteur de recherche) et mktopics (module de construction de concepts). L'interface utilisateur supporte trois modes de recherche : *plein texte* (à partir de mots-clés ou de combinaisons booléennes de mots clés) ; *par formulaire* (par filtrage sur les champs de la fiche signalétique rattachée au document) ; par *topics* ou par concepts. Un « topic » est une structure hiérarchique arborescente qui correspond à un critère de recherche complexe combinant et pondérant des opérateurs logiques, des opérateurs de proximité, des opérateurs « thesaurus » et des opérateurs qui gèrent les voisins orthographiques et phonétiques. Les feuilles terminales de l'arbre de recherche correspondent à des mots du lexique. Ce logiciel a

des avantages et des inconvénients [2], mais certaines conclusions générales peuvent être tirées qui seront valables également pour d'autres extracteurs.

Que retenir ? Quand on s'intéresse aux différents aspects d'un *thème* de discours, il est souvent utile de construire le corpus de travail avec des grains plus fins que les textes, (ex. : les paragraphes). Le tamisage de ces grains plus fins doit alors se faire par des critères de contenu sémantique qui imposent, de fait, une indexation plein texte et des procédures de scoring.

On notera surtout que ces différents fragments doivent être reliés à leur fiche signalétique (celle de leur texte d'origine). Nous ne saurions trop conseiller à nos collègues d'examiner avec soin cette question, en fonction des spécificités de leur chaîne logicielle. La question apparemment simple de du rattachement l'étiquette signalétique du texte aux fragments qui en sont extraits structure, en pratique, l'architecture de la chaîne d'extraction, traitement et archivage. Faute de l'avoir réglée en détail au début, nous avons dû refaire notre copie en raison de la taille imposante des tables engendrées par nos architectures locales. Ce qui "passait" très bien sur un court extrait de la base devenait impraticable sur la maquette complète.

Une autre question fondamentale - mal réglée par TOPIC - est celle des scores de pertinence des paragraphes, qui conditionnent l'extraction (on ne retient pour composer le corpus que les paragraphes dont le seuil de pertinence est supérieur à un certain seuil). Pour TOPIC, l'unité d'indexation et de recherche est le document physique. Cela rend difficile la granularisation du matériel textuel. Une sur-couche logicielle développée par notre partenaire TRIEL a permis de contourner cette limitation de TOPIC en lui faisant prendre les paragraphes pour des documents physiques. Cependant, la solution est lourde et mal adaptée au balisage SGML. Enfin, le mode de calcul des scores de TOPIC n'est pas transparent.

Nous n'avons pas trouvé sur le marché de solutions satisfaisantes pour scorer des paragraphes ou des fenêtres de contexte glissantes, qui serait la méthode idéale pour extraire des corpus homogènes tamisés à grain fin. Cette direction de recherche semble devoir être développée.

3. La signalétique et le stockage sous des formes gérables

Nous avons traité ces deux problèmes d'un même mouvement.

Nous avions d'abord utilisé une base de données qui reliait les étiquettes (signalétique) aux textes. Cette solution oblige à gérer deux bases (une de données, une de textes). La lourdeur d'exploitation et de mise à jour nous a amenés à une solution plus élégante. Chaque texte, étiqueté par le signalétique, constitue un document unique (texte + étiquette). Chaque document est en format SGML, et autoporteur de son signalétique, suivant une DTD unique [2]. Les documents sont empilés en vrac dans des "collections", qui ne sont rien d'autre que des fichiers sur disque, dans des dossiers du gestionnaire standard de Windows.

La DTD Scriptorium spécifie les règles de balisage des documents de la base. Elle fournit à la fois un modèle conceptuel en définissant des éléments structurels qui peuvent apparaître dans

un document, et un modèle opérationnel qui spécifie l'ordre de ces éléments et leurs occurrences dans le document (optionnel, obligatoire, répétitif). La DTD permet de définir l'unité de fragmentation (pour le moment, le paragraphe), et de lui rattacher l'information descriptive du cartouche. La DTD fournit donc un modèle de granularisation du matériel textuel et les paramètres de découpage sur lesquels vont s'appuyer les outils d'extraction pour produire des corpus thématiquement homogènes.

Cette solution est extrêmement robuste, dans la mesure où elle permet d'ajouter et d'extraire des documents à la collection sans l'utilisation d'un logiciel intermédiaire. Elle permet également de rendre la collection indépendante des logiciels d'extraction ou d'analyse (pas de format propriétaire). En même temps, le retour à un document initial permet d'obtenir immédiatement ses données signalétiques. On notera que les procédures d'extraction peuvent se faire directement avec des outils de bas niveau sur le document lui-même, dans une même boucle, puisqu'on peut examiner d'un même mouvement en recherche plein texte si le document contient (aux balises ad hoc) les éléments signalétiques, et/ou les thèmes ou mots clés recherchés. Le balisage utilisé, qui inclut la norme HTML, facilite la navigation dans le corpus et entre extractions et documents d'origine, avec les browsers usuels. Une évolution convergente a poussé nos collègues du CEMAP d'IBM vers une solution du même type.

Ces « *collections autoporteuses* », à mi chemin entre le vrac et le fichier statistique, ont l'avantage de la simplicité et de la robustesse. Ces qualités deviennent essentielles quand on manipule des gros volumes de documents (à partir d'une certaine taille, les "petits problèmes" qu'on avait l'habitude de traiter au cas par cas à la main prennent une proportion telle qu'il est impossible de faire du bricolage, il ne peut plus s'agir que de procédures globales). Nous ne lui avons pour le moment trouvé qu'un désavantage : la nécessité de tenir un compte séparé du nombre total et du type de documents contenus dans la base.. Mais sur de telles collections, de nombreuses opérations de maintenance, depuis le dédoublonnage jusqu'à l'enrichissement automatique du cartouche à partir de bases de connaissances du monde, ou le scoring, peuvent facilement être réalisées par des agents logiciels en asynchrone. Nous pensons qu'il s'agit là du format de stockage de l'avenir pour l'ADT.

4. L'historisation.

L'historisation recouvre deux problèmes différents, tous deux liés au fait que les textes produits le sont dans un certain contexte, nous les avons traités d'un même mouvement.

Le premier problème est la remise en contexte historique (« hors texte ») des textes que nous collectons. La compréhension d'un texte nécessite la connaissance du monde, ne serait-ce que parce que les objets auxquels il se réfère ne sont pas toujours explicités en détail dans le texte. C'est vrai pour une interprétation strictement linguistique, et l'est encore plus pour une interprétation sociologique comme celle que vise Scriptorium.

A l'origine, nous voulions construire une chronologie des événements, les décrivant, séparée de la base textuelle, et sur laquelle pointeraient les textes. Par exemple, la directive européenne sur le marché de l'électricité a été le sujet de nombreuses controverses qui se traduisent par des rapports, des tracts, des réunions etc. Puis nous avons réalisé que les descriptions d'événements étaient elles-mêmes des textes, voire des événements discursifs, comme la directive, ou des prises de positions publiques de tel ou tel acteur (discours du président, etc.). Nous avons donc pris la décision de considérer la base comme une série de documents qui se renvoient les uns aux autres. Ceci fait de nos collections des hypertextes à plusieurs voix, dont certaines, pas toutes, sont celles de nos acteurs internes.

Techniquement, ceci est facile à implémenter puisque nous sommes déjà en SGML. Il suffit de créer des liens entre textes, et les qualifier pour signaler par exemple des filiations référentielles, causales, ou contextuelles. Pratiquement, nous en sommes au tout début : ceci est donc une de nos directions de travail.

Le second problème est la remise en co-texte des analyses faites sur Scriptorium. Alors que la première historisation était celle du monde, la seconde, plus modeste mais aussi difficile, est celle de la base elle-même. En effet une des difficultés de la gestion de notre base est son expansion constante. Les traitements successifs sont effectués sur des extractions d'une masse variable. Ainsi, une même requête d'extraction donnera deux résultats différents d'une année sur l'autre, même si elle porte sur la même période historique « hors-texte » (ex : 1960-1995).

Pourquoi ? Car l'ordre d'entrée des documents dans la base ne correspond pas forcément à leur ordre de production. On peut ainsi insérer en 1996 des articles de journaux internes écrits en 1995, et en 1997 des textes écrits en 1964, selon les bonnes fortunes de la moisson documentaire. Par exemple, on a obtenu, trois ans après, l’accès aux réponses d’un échantillon supplémentaire de la vague d'une enquête dont on avait déjà traité les questions ouvertes. Cette particularité devient un problème quand on veut réaliser des études longitudinales, ce qui est précisément un des objectifs principaux de Scriptorium, dédié à des analyses de fond sur longue période. Il faut donc archiver avec chaque analyse des informations sur le corpus utilisé. Réciproquement, chaque analyse apporte éventuellement des informations supplémentaires sur chaque texte. Pour les exploiter, nous envisageons des méthodes permettant l'enrichissement progressif des documents par du méta-texte. Ceci revient à enrichir le cartouche.

Une perspective plus générale de l'historisation de la base nous amène donc à considérer celle-ci comme une collection croissante de documents, qui peuvent être annotés, et d'analyses tirées de ces documents, qui sont des documents eux-mêmes. De fait, des séries de document ou un document particulier peuvent servir de contexte aux autres. C'est notamment le cas lorsque des controverses entre acteurs produisent des documents qui se répondent. Ces indications peuvent servir à structurer des coupes sur un cours d'événements donné.

Cette structuration en hypertexte des collections nous paraît la seule voie praticable pour une historisation correcte. Techniquement, les balises de la DTD Scriptorium actuelles peuvent être intégrées en tant que sur-couche à la DTD ISO 12083 (celle de HTML). Le comportement d'un browser vis-à-vis des balises non définies dans la norme est simplement de les ignorer. Cette structuration en hypertexte pose cependant, comme on l'a évoqué, divers problèmes théoriques que nous n'avons pas tous résolus.

5. La construction de concepts

Un dernier aspect du traitement d'une telle base historique nous pose des difficultés toutes spéciales. Rappelons que nous travaillons dans une perspective exploratoire, et que nous cherchons à décrire des évolutions dans le temps et l'espace social. Par exemple, nous repérons, à l'aide d'Alceste, les thèmes abordés par tel ou tel acteur. Cette analyse nous donne, en l'état de l'art des résultats satisfaisants [1, 5, 6] ; mais comment examiner l'évolution, sur une base comparable, de l'importance de tel aspect du thème (par exemple : l'aspect « relation au client » dans le thème du service public) quand le corpus aura grossi ? Une nouvelle classification des thèmes modifie, comme nous en avons fait l'expérience, à la fois *la taille* et *le contour* des classes. Ces modifications sont en général marginales, mais comment faire le départ entre l'évolution en volume et l'évolution en nature d'une classe thématique ?

Pour cela, nous sommes amenés à figer des concepts, qui servent d'état de référence, et permettent un scoring des différents textes ou fragments sur des bases comparables. Pour le moment, ces concepts sont des « concepts » Topic, c'est à dire des combinaisons arborées d'opérateurs sur des items lexicaux (en fait, un « concept » ressemble fort à une requête sophistiquée sur une base documentaire). Nous faisons actuellement des essais pour construire directement ces concepts à partir des sorties d'Alceste, notamment des listes de traits typiques et des chi deux de liaison de ces traits typiques avec les classes (ce qui donne de grands arbres pondérés).

De tels concepts, qui restent stables dans le temps - aux évolutions près de la signification des mots dans la langue -, permettent également de scorer directement des textes nouveaux, et d'enrichir en sur-texte nos collections ou des fichiers qui en sont issus et destinés à des recherches statistiques plus sophistiquées.

Nous ne sommes qu'aux premiers pas de cette démarche, et les solutions techniques adoptées sont provisoires, mais nous avons la conviction que de tels outils de formalisation du contenu des thèmes seront nécessaires. Les listes de vocabulaire construites manuellement, dans des logiciels comme Tropes (« Univers de référence »...) ou Prospero (« catégories »...), qui sont centrés sur l'analyse thématique ; les « clés » d'Alceste, etc. vont dans ces directions ; mais il nous paraît nécessaire de disposer de méthodes pour construire de tels objets directement à partir des résultats d'analyse, pour pouvoir opérer plus rigoureusement des comparaisons de contenu. C'est, là encore, une direction qui devra certainement être creusée.

6. Conclusion

Le traitement de grandes bases de données de textes hétérogènes par ADT est certainement une direction d'avenir ; elle nécessite des aménagements par rapport aux stratégies actuelles d'exploitation des corpus homogènes.

Nos premiers pas dans une telle base de données textuelles constituée en vue d'une exploitation systématique par ADT nous a montré, d'abord, que le volume apporte des difficultés en soi. Nombre de petits problèmes qui se réglaient à la main, ad hoc, appellent une solution systématique sur gros volume. Ensuite, à ces échelles, la mobilisation d'outils spécialisés avec leurs formats spécifiques, notamment les SGBD, nous a paru trop lourde et rigide.

Une autre difficulté de l'exploitation de gros volumes est la nécessité de tamisage à grain fin à base d'indexation plein texte. Cette difficulté, négligeable pour ceux qui travaillent à partir d'une veine homogène en conditions de production (par exemple des extraits d'un CD-ROM de journal quotidien) devient cruciale pour du travail sur minerai textuel hétérogène.

Nous proposons pour la gestion des textes destinés à l'ADT un format particulier, les *collections autoporteuses*, où chaque document est «autoporteur» de son signalétique, et où d'autres renvois au contexte mondain sont faites par hyperliens. Dans l'état actuel de la technique, le balisage SGML et les hyperliens permettent de réaliser sans trop de difficultés de telles collections. L'avantage du balisage SGML est qu'il peut également fournir un prédécoupage du texte en unités à grain fin, comme les paragraphes.

L'évolution de nos collections de textes nous amène à poser non seulement les questions d'historisation des textes (mondaine et technique), mais aussi celle de la formalisation des résultats des analyses sous des formes figées qui permettent des comparaisons. Nous testons actuellement des solutions à base de « topics » structurées comme des requêtes d'indexation.

REFERENCES

[1] Folch, H., Lemoine, J-C., Lahlou, S. (1996). *Scriptorium, premiers essais de la maquette.* HN-51/96/017. EDF-DER.

[2] Folch, H. (1997). *Bilan technique de la plate-forme SCRIPTORIUM - août 1997* -. HN-51/97/ 015. EDF-DER.

[3] Lahlou, S, Piat, G., Aubert, C. (1995). *Scriptorium : le projet*. HN-51/95/007. EDF-DER.

[4] Lahlou, S. (1996) La modélisation de représentations sociales à partir de l'analyse d'un corpus de définitions. In : Martin E. (éd.). *Informatique textuelle*. Coll. Etudes de Sémantique Lexicale. INaLF. Paris : Didier Erudition. pp. 55-98.

[5] Lemoine, J-C. (1996) *Test de fonctionnement de l'outil Scriptorium sur un petit corpus.* HN-51/96/023. EDF-DER.

[6] Lemoine, J-C. (1997). VVE 1995. *Utilisation de Scriptorium pour l'analyse de discours syndicaux et directoriaux a EDF-GDF sur la période 1991-1995.* HN-51/97/ 016. EDF-DER.

[7] Reinert, M. ALCESTE, une méthode d'analyse des données textuelles. Application au texte "Aurélia" de Gérard de Nerval. *B.M.S.,* 26, 1990: 25-54.