Pré-analyse de corpus

Thierry Poibeau

Laboratoire d'Informatique de Paris Nord (LIPN)
Université Paris 13 – Av. J-B. Clément – 93430 Villetaneuse – France
poibeau@lipn.univ-paris13.fr

Abstract

Most Natural Language Processing tools need homogeneous corpora in order to deliver relevant results. However, such corpora are rarely available in industrial and applicative contexts. This paper presents an original approach for preparing corpora in order to obtain useful amount of texts. The presented techniques are based on statistical and surface linguistic analysis. We present these techniques and an experiment in the information extraction domain. We demonstrate the different techniques and evaluate their interest for the task.

Résumé

Les applications de Traitement Automatique des Langues nécessitent le plus souvent des corpus homogènes pour fournir des résultats pertinents. De tels corpus sont rarement disponibles dans des contextes applicatifs ou industriels. Cet article propose une approche originale pour préparer les corpus et obtenir des masses de textes utilisables. Les techniques présentées vont des statistiques à l'analyse linguistique de surface. Les techniques sont présentées puis appliquées au domaine de l'extraction d'information. Nous montrons l'intérêt de ces techniques et nous en donnons une évaluation fondée sur leur pertinence pour la tâche visée.

Mots-clés : pré-analyse de corpus, analyse linguistique de surface, approches statistiques

1. Introduction : analyse et représentation du contenu des textes

Le Traitement Automatique des Langues s'est beaucoup intéressé dans les décennies passées à la compréhension de textes. La compréhension, au moins partielle, est nécessaire pour le filtrage, l'aide à la décision et la traduction. Devant la difficulté de la tâche, une approche fondée sur une analyse partielle a été proposée, à travers notamment les techniques d'extraction d'information. Il ne s'agit plus de comprendre tout le texte mais d'extraire un certain nombre d'informations pertinentes par rapport à un formulaire prédéfini en fonction de la tâche visée.

Même si l'extraction est actuellement une technologie bien maîtrisée, la question de la mise au point des systèmes reste un problème crucial. Plusieurs auteurs ont proposé le recours à l'apprentissage pour l'acquisition semi-automatique de lexiques et de grammaires d'extraction mais ces techniques laissent de côté plusieurs aspects importants : comment savoir si un ensemble de textes se prête bien à une analyse en vue d'une application d'extraction d'information ? L'ensemble de textes considéré porte-t-il sur des faits avérés ou s'agit-il de déclarations ou de rumeurs ? Quelle est la part de modalité ou de négation, phénomènes que l'on sait mal analyser, surtout sur du texte « tout venant » ? Ces phénomènes posent des problèmes majeurs aux stratégies d'analyse locale telles qu'elles sont mises en œuvre dans le cadre des applications d'extraction.

Dans de nombreux cas, les concepteurs d'applications partent de corpus hétérogènes faits de plusieurs types de textes traitant de différents domaines. Or, on sait que l'extraction nécessite des corpus homogènes et qu'il faut au préalable élaborer un ensemble de textes pertinents pour pouvoir l'analyser puis en extraire des connaissances. Il existe heureusement des outils permettant de procéder à une pré-analyse de corpus afin d'en déterminer les principales caractéristiques. Les techniques de cartographies permettent notamment d'opposer plusieurs sous-corpus d'après certains traits caractéristiques. Cet article montre une utilisation d'outils d'analyse de corpus – statistiques et linguistiques – dans cette perspective applicative.

Nous présentons tout d'abord les techniques à l'œuvre. Nous détaillons ensuite les différentes expériences effectuées et nous montrons ainsi la complémentarité entre les approches et les outils utilisés pour cette tâche. Nous procédons ensuite à une évaluation des différentes techniques mises en œuvre.

2. Techniques de pré-analyse de corpus

Les techniques mises en œuvre pour la pré-analyse de texte sont surfaciques, pour des raisons évidentes d'efficacité. On trouve essentiellement des méthodes statistiques mais aussi quelques analyses de type linguistique.

2.1. *Méthodes quantitatives et analyse de spécificités*

Les méthodes statistiques permettent des traitements rapides sur de gros corpus. Ce type de méthode peut ainsi être valablement utilisé pour la partition d'un corpus pour peu que l'on dispose d'une base d'entraînement. Dans le cadre qui nous occupe (pré-analyse de textes dans la perspective d'applications d'extraction d'information) cette condition est généralement remplie. Les analystes sont le plus souvent des experts ayant emmagasiné des textes pertinents dans leur domaine de spécialité.

L'analyse de spécificités peut porter sur les formes du texte, sur des formes lemmatisées ou sur une analyse de spécificités. Nous utilisons l'outil LEXICO d'A. Salem *et al.* (2002), qui permet de combiner analyse de spécificités et recherche de segments répétés (« suite de formes dont la fréquence est supérieure ou égale à deux dans le corpus », d'après (Lebart et Salem, 1994)). Ces formes sont souvent plus représentatives d'un corpus que les formes simples. On peut alors procéder à un calcul de spécificités positives et négatives : « pour un seuil de spécificité fixé, une forme *i* et une partie *j* données, la forme *i* est dite spécifique positive (resp. négative) de la partie *j* (ou forme caractéristique de cette partie) (...) si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures (resp. inférieures) à la sous-fréquence constatée est inférieure au seuil fixé au départ. » (repris de (Lebart et Salem, 1994)).

Une analyse factorielle de correspondance (AFC) permet de projeter les résultats sur un plan et de faire apparaître certains regroupements. Si ce type d'étude donne souvent des résultats difficiles à interpréter, nous montrons dans ce qui suit que l'AFC peut permettre de dégager automatiquement des regroupements de textes intéressants au sein d'un corpus hétérogène. Nous utilisons cette technique de manière très simple pour faire apparaître des sous-corpus au sein d'un ensemble hétérogène de textes.

2.2. *Analyse linguistique de surface*

L'analyse linguistique vise à préciser et à affiner les résultats de l'analyse statistique. Nous avons vu dans l'introduction que tous les textes ne se prêtent pas bien à l'extraction d'information. Si le texte comporte trop de modalités, si l'analyse linguistique risque d'être

troublée par des phénomènes de portée (de quantifieurs, de négations, etc.), il importe de se montrer prudent sur la valeur du résultat obtenu. On sait en effet qu'il s'agit là de marqueurs de prise de distance par rapport au fait, pouvant introduire un doute ou nuancer l'information rapportée. L'analyse de spécificités de textes non événementiels (textes qui ne relatent pas que des faits mais qui incluent aussi des commentaires et des prises de position) met souvent en avant des formes modales et des négations qui doivent être prises en compte pour juger de la pertinence globale que l'on cherche à atteindre. Si l'analyse de spécificités montre trop de formes modales, alors l'analyse doit être regardée avec prudence et un retour au texte peut s'avérer nécessaire.

Enfin, une fois que l'on a obtenu un corpus homogène et pertinent, il peut être utile de filtrer les sous-parties du corpus pertinentes pour la tâche visées. L'expérience acquise dans le cadre des conférences américaines d'évaluation MUC a montré que, souvent, moins de 10 % du corpus était utile (MUC, 1995). Dans certains cas extrêmes comme la recherche de relations d'interaction au sein d'un corpus de génomique, environ 3 % du texte est pertinent. Des techniques simples et surfaciques peuvent permettre d'effectuer ce filtrage (Bessières *et al.*, 2001). Nous avons précédemment expérimenté une technique fondée sur la densité sémantique dans (Poibeau, 2003) : un filtrage thématique peut aisément être implanté au moyen de mots clés (généralement fournis par l'utilisateur) enrichis sémantiquement par le recours à un réseau de type Wordnet. Nous montrons ici comment l'analyse de formes particulières comme les entités nommées peut également favoriser ce filtrage.

3. Expériences

Nous avons constitué un corpus composé pour une moitié de dépêches AFP et pour une autre moitié de courriers électroniques. Afin d'envisager un traitement automatique, il est nécessaire de pouvoir partitionner cet ensemble de textes en deux sous-ensembles distincts. Pour ce faire, nous utilisons le logiciel LEXICO décrit dans la partie 2.1.

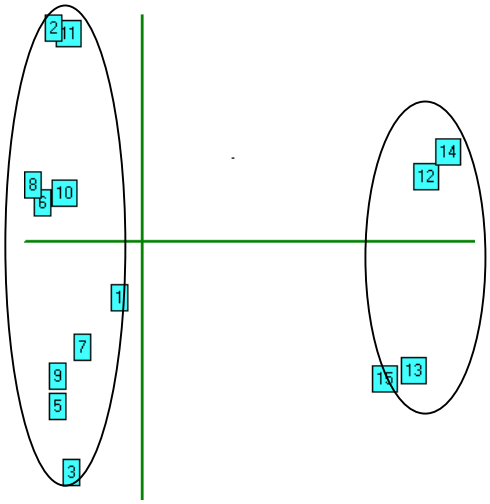

Figure 1. Caractérisation de corpus par analyse statistique.
On voit clairement apparaître deux partitions, entourées d'un trait noir sur le graphique.

Nous donnons ci-dessus une carte produite par analyse factorielle de correspondance produite suite à une analyse de spécificités du corpus. Le résultat fait clairement apparaître deux ensemble de points, répartis de part et d'autre d'un axe vertical. Un retour au texte permet de voir que l'ensemble de gauche correspond aux courriers électroniques et que l'ensemble de droite correspond aux dépêches AFP. L'analyse ne prend pas tant en compte des oppositions d'ordre lexical que des oppositions d'ordre morpho-syntaxique, voire typographique. Il est

aussi possible d'opposer les textes d'un même domaine mais appartenant à des genres différents (cf. Biber, 1995 et 1998).

L'analyse de la carte obtenue est révélatrice du contenu des documents et de la façon dont ils se prêtent à l'extraction d'information. L'ensemble de points situé à gauche (les courriers électroniques) est marqué par un ensemble de facteurs peu propices à l'extraction : présence de pronoms de la première personne du singulier, d'opérateurs de modalités, etc. Quand on analyse l'ensemble des dépêches AFP au regard de ce corpus, les spécificités négatives (formes sous-représentées dans le corpus AFP par rapport au corpus de courriers électroniques) sont à cet égard très éclairantes :

Terme	Fréquence	Spécificité
Je	103	-10
Windows	47	-9
ai	41	-9
J'	29	-9
Comment	53	-8
dois	23	-8
un	122	-7
puis	18	-6
mon	21	-6
insérer	14	-6

Figure 2. Analyse de spécificités. On voit nettement apparaître que les formes qui pourraient laisser planer un doute sur la faisabilité de l'extraction (présence de verbes à la première personne, de modaux, ...) apparaissent dans la série des formes sous-représentées. Il s'agit donc d'un corpus pertinent pour une application d'extraction.

Il est ainsi possible, en mettant en œuvre des moyens simples et automatiques, de caractériser grossièrement des corpus afin de déterminer leur adéquation à une tâche donnée. Il serait sans doute possible d'aller plus loin et de prédéfinir une série d'expressions ou de marqueurs se prêtant mal à une application d'extraction. Cette stratégie est proche de l'*exploration contextuelle* définie par ailleurs par Minel (2003). Le système ne peut bien évidemment pas prendre de décision automatique, le choix final revenant à l'utilisateur. C'est à lui de juger de la pertinence du texte, de voir si une application d'extraction est envisageable[1] mais une série de marqueurs pourrait quand même être définie *a priori*.

L'extraction d'information est largement fondée sur l'analyse de faits impliquant des entités nommées. De fait, les passages impliquant des entités nommées sont souvent pertinents pour les tâches d'extraction. Il existe aujourd'hui plusieurs systèmes permettant de repérer de manière relativement pertinente des entités nommées. Des techniques d'analyse de densité et de visualisation de données peuvent ensuite faire apparaître de manière claire les passages

[1] Cette tâche suppose un niveau d'expertise technique qui laisse supposer que des utilisateurs experts ou des administrateurs ayant à la fois des connaissances sur le domaine technique visé, et des connaissances en modélisation linguistique. Il semble encore illusoire de vouloir donner les outils de conceptions d'outils linguistiques directement aux utilisateurs finaux, sauf pour des tâches limitées.

marqués par une forte concentration d'entités. La figure ci-dessous montre une telle analyse sur une partie de notre corpus.

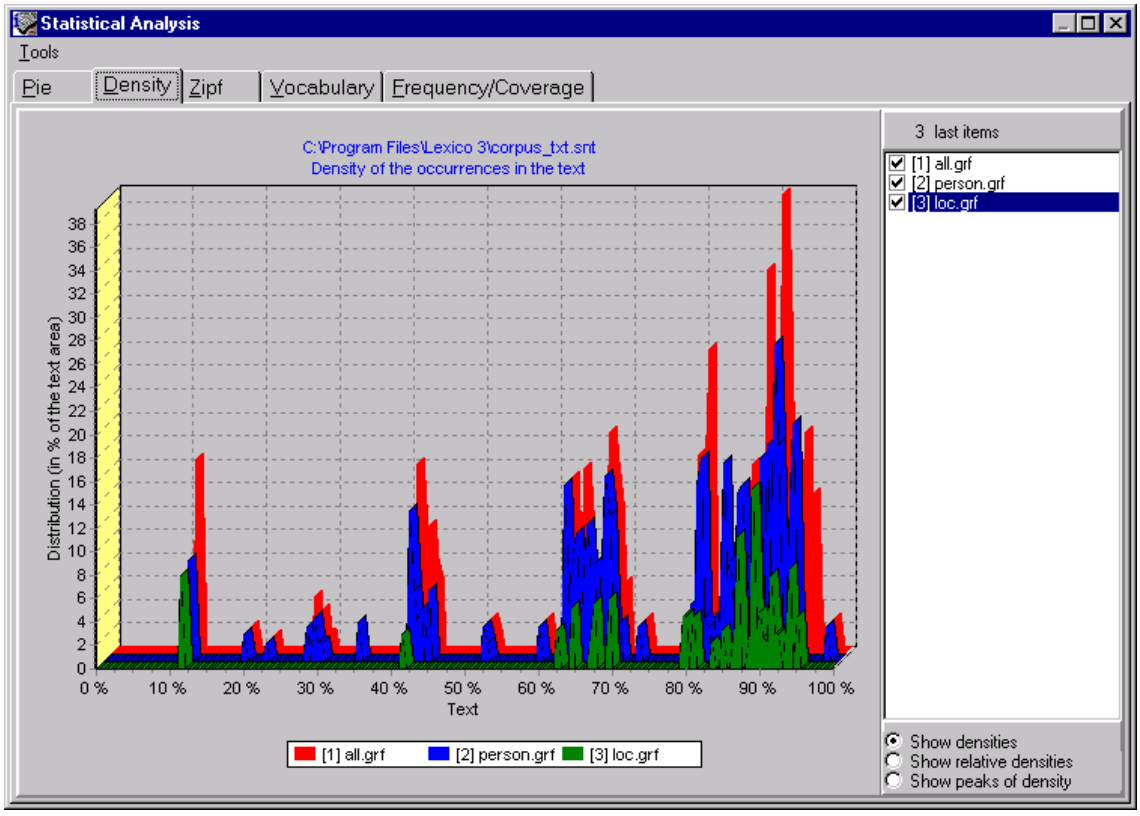

Figure 3. Analyse de densité portant sur les entités nommées du corpus. Une première étape a permis de marquer au sein du corpus les entités reconnues. L'outil statistique ici présenté (module statistique d'Intex, cf. Silberztein 1993) permet de repérer les zones marquées par une forte densité d'entités.

La figure 3 présente des courbes de densité d'entités nommées pour le corpus étudié. La courbe en supérieure rassemble le résultat de l'ensemble des entités, les deux courbes inférieures révèle pour l'une la présence de noms de personnes et pour l'autre la présence de noms de lieux. Les trois courbes sont largement homogènes sur tout le corpus.

4. Évaluation

La mise au point du système décrit dans la section précédente a porté sur un ensemble réduit de textes. Elle a été reproduite sur un corpus plus volumineux afin d'obtenir des chiffres plus fiables.

4.1. *Le corpus*

Un ensemble de dépêches provenant de la société FirstInvest a été utilisé comme corpus de référence. Il s'agit d'un « corpus de suivi » (Habert *et al.*, 1997) dans la mesure où l'on a affaire à un amas de textes qui grossit chaque jour, par opposition aux corpus généralement utilisés en linguistique, qui sont constitués une fois pour toutes sur un thème donné.

Afin d'établir un corpus de référence, un ensemble de 2000 dépêches réparties en 26 thèmes différents nous a été fourni. Nous avons extrait de ce corpus les dépêches se rapportant au thème des rachats d'entreprises (acquisitions, cessions, prises de participation, offres publiques d'achat, …). Le corpus obtenu à partir du thème « rachat d'entreprises » est composé de

303 dépêches (soit 56 000 mots). Ce corpus a été analysé manuellement pour pouvoir procéder à une évaluation de type MUC mais ici nous nous contentons du corpus brut, qui est suffisant pour nos expériences. Ce corpus est divisé en deux : un ensemble de 243 textes a constitué le corpus d'entraînement (45 334 mots), les 60 textes restants formant le corpus de test (10 666 mots).

4.2. *Évaluation*

Le corpus correspondant au thème « rachat d'entreprise » a été mêlé à un ensemble de textes de taille correspondante sur le thème « rumeur » (environ 15 000 mots) puis sur le thème « nouvelle économie » (environ 24 000 mots). Les résultats sont contrastés. Nous avons procédé à deux expériences, en mêlant tout d'abord les textes du thème « rachat d'entreprises » à ceux du thème « rumeur » puis au thème « nouvelle économie ». Nous partitionnons ensuite le corpus en deux en utilisant Lexico comme précisé ci-dessus. Il est alors possible de mesurer le nombre de textes bien catégorisés (c'est-à-dire, correspondant à la classe initiale). Nous obtenons les résultats suivants :

	Bien classés	Mal classés
Thème « rachat d'entreprise »	91 %	9 %
Thème « rumeur »	79 %	21 %

	Bien classés	Mal classés
Thème « rachat d'entreprise »	82 %	18 %
Thème « nouvelle économie »	77 %	23 %

On voit sur ces résultats que les textes sont globalement bien classés, même si des erreurs subsistent. Le premier constat est que le système de classement ainsi constitué fonctionne mieux sur les thèmes « rachat d'entreprise » *versus* « Rumeur » que sur « rachat d'entreprise » *versus* « Nouvelle économie ». Comme on peut aisément l'imaginer, les textes du thème « Rumeur » comportent en effet des modaux et des adverbes de prises de distance, qui sont à l'inverse peu présent dans le thème « rachat d'entreprise ». Les termes employés dans le corpus « nouvelle économie » sont en revanche moins discriminants. On obtient donc des résultats moins bons : 23 % des messages se rapportant à ce corpus étant classés dans le thème « rachat d'entreprise » dans notre expérience. Il faut cependant relativiser ces résultats : un texte peut appartenir à un thème donné et être aussi pertinent pour un autre thème.

Nous avons ensuite testé la sélection de passages à partir des entités nommées. Un filtrage uniquement conçu sur les entités nommées n'est pas performant si l'on compare les résultats obtenus à partir de cette méthode avec les résultats d'une méthode fondée sur un ensemble de mots clés. Des passages pertinents ayant une faible concentration d'entités nommées ne sont pas repérés et, à l'inverse, les entités ne semblent pas être un critère suffisant pour retenir un passage. Une étude manuelle des résultats montre cependant que les passages comportant des entités nommées sont plus pertinents que les passages sans entités nommées. Il est donc possible d'éliminer automatiquement un certain nombre de passages non pertinents. La proportion de texte retenu peut ainsi passer de 20 % à 15 % du texte original sans conséquence majeure pour la tâche suivante (la phase d'extraction proprement dite). On a ainsi un gain de temps de traitement du fait de l'amélioration du filtrage initial marqué par un nombre infé-

rieur de textes à traiter (passage de 20 à 15 % de la masse textuelle initiale, soit une diminution de 25 %).

5. Conclusion

Nous avons présenté dans cet article quelques expériences utilisant des méthodes rapides afin de caractériser *a priori* un corpus dans la perspective d'une application d'extraction d'information. Nous avons ainsi montré qu'il était aisé d'obtenir une analyse statistique (à partir d'outils comme Lexico) permettant de partitionner un corpus afin d'obtenir un ensemble de textes homogène. De plus certains indices linguistiques (modaux, négation, etc.) permettent de prédire de façon sommaire mais efficace si les textes en question peuvent se prêter à une application d'extraction. Dans certaines conditions, la pré-analyse des textes peut inciter à être prudent sur les résultats et à les valider manuellement. Nous avons enfin présenté un filtrage fin des données linguistiques en fonction de l'analyse des entités nommées. Les résultats que nous avons présentés offrent une analyse fine des phénomènes en jeu. Ce travail reste à compléter afin d'obtenir des systèmes d'extraction ou de compréhension opérationnels sur des données réelles.

Références

Bessières P., Nazarenko A. et Nédellec C. (2001). Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. In *4e Colloque International sur le Document Électronique* (CIDE'201) : 165-183.

Biber D. (1998). *Variation across speech and writing*. Cambridge University Press.

Biber D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press.

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Armand Colin.

Lafon P. (1984). Pour une présentation de la méthode et du calcul de la probabilité. In *Dépouillements et statistiques en lexicométrie*. Champion-Slatkine.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Minel J.-L. (2003). *Filtrage d'information*. Hermès.

MUC (1995). *Proceedings of the 6th Message Understanding Conference (MUC 6)*. Morgan Kaufmann.

Poibeau T. (2003). *Extraction d'information, du texte brut au web sémantique*. Hermès.

A. Salem *et al.* (2002). *Manuel d'utilisation de Lexico.* Université Paris 3.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes, le système INTEX*. Masson.