Acquisition automatique de terminologie à partir de corpus de texte

Edmond Lassalle

(1) Orange Labs, 2 avenue Pierre Marzin
22 307 Lannion - France
edmond.lassalle@orange-ftgroup.com

Résumé :

Les applications de recherche d'informations chez Orange sont confrontées à des flux importants de données textuelles, recouvrant des domaines larges et évoluant très rapidement. Un des problèmes à résoudre est de pouvoir analyser très rapidement ces flux, à un niveau élevé de qualité. Le recours à un modèle d'analyse sémantique, comme solution, n'est viable qu'en s'appuyant sur l'apprentissage automatique pour construire des grandes bases de connaissances dédiées à chaque application. L'extraction terminologique décrite dans cet article est un composant amont de ce dispositif d'apprentissage. Des nouvelles méthodes d'acquisition, basée sur un modèle hybride (analyse par grammaires de chunking et analyse statistique à deux niveaux), ont été développées pour répondre aux contraintes de performance et de qualité.

Abstract :

Information retrieval applications by Orange must process tremendous textual dataflows which cover large domains and evolve rapidly. One problem to solve is to analyze these dataflows very quickly, with a high quality level. Having a semantic analysis model as a solution is reliable only if unsupervised learning is used to build large knowledge databases dedicated to each application. The terminology extraction described in this paper is a prior component of the learning architecture. New acquisition methods, based on hybrid model (chunking analysis coupled with two-level statistical analysis) have been developed to meet the constraints of both performance and quality.

Mots-clés : Apprentissage automatique, acquisition terminologique, entropie, grammaires de chunking
Keywords: Unsupervised learning, terminology acquisition, entropy, chunking analysis

1 Introduction

Une amélioration significative de la qualité des moteurs de recherche concerne l'identification des locutions en tant qu'unités de sens. C'est aussi une difficulté dans le cas de certaines applications d'Orange. Le problème est en effet de pouvoir prendre en compte une terminologie en constante évolution dans des domaines liés à l'actualité (presse, journaux télévisés...). Il s'agit en plus de traiter en continu des flux importants de données pour indexer les nouveaux documents entrants mais aussi pour acquérir une terminologie évanescente (*fuite de pétrole*, *nuage de cendres*, *Jean Paul II*, *Sidi Bouzid*, *Antoine de Léocour* ...). Les méthodes d'acquisition automatique de terminologie à partir de corpus trouvent ici leur entière justification.

Un examen de différents modèles d'apprentissage, de leur adéquation aux corpus dans nos applications va motiver une architecture hybride différente de celles connues et étudiées à ce jour. Ce choix oblige à innover dans les méthodes d'analyse linguistique et statistique pour répondre aux contraintes opérationnelles de qualité. L'objet de nos travaux est alors, d'avoir un système «homogène» pour limiter le biais statistique inhérent aux interactions dans tout modèle hybride. La loi binomiale régissant le comportement des mots constitue donc la seule hypothèse de départ. Des observations expérimentales, une modélisation formalisée permettent ensuite de dériver par calcul les autres lois. Les résultats obtenus vont confirmer la pertinence de cette démarche. Dans la suite de l'article, une description du modèle d'apprentissage, des méthodes d'analyse statistique va donner un éclairage sur le fonctionnement de notre composant linguistique.

2 Motivation d'un modèle hybride d'acquisition terminologique

Le choix d'une architecture est dicté par le type de corpus d'apprentissage. Le nôtre est constitué de textes décrivant des vidéos sur un mois d'actualités (http://www.2424actu.fr/actualite-du-jour/). A chaque instant, on dispose de 100 000 textes pour un total de 5 millions de mots. Chaque texte comprend un titre suivi d'un résumé court comme : «*Tunisie : affrontements à **Sidi Bouzid**. De nouveaux affrontements violents ont eu lieu dans la nuit dans la région de **Sidi Bouzid**, dans le **centre-ouest** de la Tunisie, faisant un blessé par balle et des **dégâts matériels** importants, a-t-on appris dimanche de **sources syndicales** Des centaines de Tunisiens ont participé à une manifestation.*»

Dans ce type de corpus, certaines locutions – étant communes (*dégâts matériels*, *sources syndicales*) – peuvent être obtenues hors méthodes d'apprentissage, mais d'autres (*Sidi Bouzid* ou *Camp Nou*) risquent de ne pas figurer dans un référentiel lexical qui serait établi *a priori*. Le problème à traiter est donc d'avoir un référentiel de mots simples exhaustif, incluant des mots inconnus. Une analyse visant à extraire des locutions devra ensuite identifier des constructions bien formées de groupes de mots, puis reconnaître la nature compositionnelle ou figée du sens porté par ces constructions, y compris celles comportant des mots inconnus. Les solutions à cette problématique peuvent être d'ordre statistique ou mixte, mais excluent une approche symbolique confrontée au problème d'exhaustivité.

2.1 Modèles statistiques

L'apport des méthodes statistiques concerne la quantification de la compositionalité. L'occurrence d'un mot m_i dans un corpus est modélisé par une loi de Bernoulli de paramètre p_i. Le comportement d'un mot dans le corpus est ensuite expliqué par sa fréquence d'occurrences et donc par une v.a.r de loi binomiale $B(n,p_i)$. Estimer le degré de compositionalité de deux mots contigus revient alors à déterminer le degré de dépendance des v.a.r associées à ces mots. Deux méthodes expérimentales permettent de réaliser ce calcul :

- La première nécessite une fenêtre d'observation (par exemple la phrase) pour estimer les probabilités d'occurrences et de cooccurrences à partir d'un comptage fréquentiel. Elle conduit au calcul de l'information mutuelle (Church et al., 1990) ou à la mesure de Dice (Smadja, 1993). Citons aussi pour cette méthode, le calcul de la log-perplexité (Kit, 2002) qui a l'avantage de prendre en compte des séquences de N mots mais nécessite en contre partie un modèle de langue pour viabiliser l'estimation de la probabilité de telles séquences.
- La seconde réalise un comptage fréquentiel direct de la cooccurrence, de la non-cooccurrence et des non-occurrences de deux mots contigus pour déterminer la log-vraisemblance des 2 v.a.r associées (Dunning, 1993) ou aussi leur corrélation via le calcul du χ^2.

Le résultat pour ces 2 méthodes est un classement suivant une «vraisemblance d'être une locution». La difficulté restant est de déterminer la valeur de seuillage, mais aussi de mesurer l'importance des termes par rapport au corpus applicatif.

Pour traiter ce dernier point, les modèles les plus avancés (Kit, 2002) (Vu et al., 2008) (Kageura et al., 1996) caractérisent les séquences extraites par le critère d'unithood, validant statistiquement la cohérence de la séquence, et par le critère de termhood, caractérisant la spécificité de la séquence par rapport au corpus applicatif. En l'absence d'analyse linguistique, le premier critère permet de valider la construction syntaxique de la séquence tandis que le second critère valide à la fois la non-compositionalité et l'importance de cette séquence. Cette approche est adaptée pour les domaines techniques où le vocabulaire est limité, où les expressions figées peuvent être longues comme *Altération des facteurs de coagulation sanguine*, où le critère de spécificité est assez proche du critère de non-compositionalité. Une variante intéressante (Frantzi et al., 1999) est d'introduire le filtrage de catégories grammaticales et de palier l'absence d'analyse syntaxique par des mesures statistiques (AC/NC-value).

2.2 Modèles hybrides

Le modèle le plus usité est basé sur un fonctionnement en tandem du composant linguistique et du composant statistique. L'avantage d'une telle architecture concerne la modularité. L'analyse linguistique est chargée d'annoter le corpus initial (étiquetage grammatical, parenthésage et étiquetage des syntagmes). L'analyse statistique reprend les informations annotées pour produire une liste de termes classés suivant un ordre de vraisemblance. Cette approche permet en plus de reprendre pour le deuxième composant (Daille, 1996) les mesures utilisées par les modèles statistiques. L'inconvénient du modèle en tandem concerne le biais statistique. Les évaluations que nous avons menées (Lassalle et al., 2011) sur Acabit ont indiqué un différentiel de 30% du taux de précision suivant que nous utilisons en amont, comme composant linguistique, l'analyseur de Brill (Brill, 1992) couplé au lemmatiseur Flem (Namer, 2000) ou l'analyseur Tilt (Heinecke et al., 2008).

Seul un couplage fin entre analyse linguistique et analyse statistique permettrait de minimiser ce biais. Ce qui exclut une réutilisation des mesures de classement des modèles statistiques car ces dernières nécessitent, dans le calcul, des données globales et non partielles comme c'est le cas dans un couplage fin. Cela nous conduit à spécialiser nos méthodes d'analyse statistique dans deux directions :

- la première pour détecter les éléments saillants (analyse de régularité)
- la seconde pour estimer la non-compositionalité des constructions syntaxiques.

Le rôle de l'analyse linguistique dans cette approche hybride est de proposer successivement des ensembles «statistiquement cohérents» de constructions syntaxiques. Ce que nous préciserons après avoir décrit dans un premier temps les analyses statistiques.

3 Analyse statistique de la régularité

Les finalités de l'analyse statistique décrite dans cette section sont triples. La même observation expérimentale permet en effet de déduire les caractéristiques des mots dans le corpus, suivant :

- une loi de distribution décrivant leur occurrence,
- des propriétés macroscopiques autorisant leur regroupement au sein de catégories grossières,
- et le degré de saillance permettant d'identifier les mots importants dans le corpus.

Seul, le calcul de saillance est prééminent dans l'acquisition de terminologie. La loi de distribution permet de déduire la loi conjuguée *a priori* et elle est plutôt utilisée dans nos modélisations bayésiennes, comme dans la catégorisation[1] ou dans l'indexation. Le regroupement des mots en catégories grossières, bien qu'utile dans le processus d'acquisition terminologique, nécessite une extension (restant à faire) du calcul de saillance.

3.1 Loi de distribution des mots

Si l'on accepte que l'occurrence d'un mot dans un corpus suit une loi de Bernoulli de paramètre p, alors sa fréquence d'apparition dans une fenêtre de n mots d'un corpus suit la loi binomiale B(n,p). La valeur de p est en général très faible, à l'exception des mots grammaticaux et des termes de domaine (dans le cas de corpus

[1] Blei D.M., (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*

spécialisés comme ceux de la médecine, des finances,...). Il est donc possible pour les grandes valeurs de n d'approximer la loi binomiale B(n,p) par une loi de Poisson ou par une gaussienne discrétisée (Saporta 2006).

L'intérêt d'une loi de Poisson P(λ) par rapport à une gaussienne est d'avoir l'espérance et la variance égales à λ. Pour les grandes valeurs de λ (λ>18), P(λ) peut être confondue à une loi de Gauss (Saporta, 2006), avec l'avantage d'être caractérisée par un seul paramètre. L'estimation d'un seul paramètre (espérance = variance) plutôt que 2 présente un gain important en qualité dans l'apprentissage à condition que la loi de Poisson soit justifiée.

Le problème est donc de savoir, à partir d'observations expérimentales, quand représenter les fréquences d'occurrence par une loi de Poisson, c'est à dire, pour les grandes valeurs de fréquence quand représenter par une gaussienne à un seul paramètre ou par une gaussienne à 2 paramètres. Nous nous appuierons sur le théorème suivant (Saporta, 2006) pour affecter expérimentalement les mots observés dans l'une de ces 2 catégories.

Théorème :
Si X_n est une suite de variables binomiales B(n,p) telles que quand $n\to\infty$ et $p\to0$, np tend vers une limite finie λ. Alors X_n converge en loi vers une variable de Poisson P(λ)

3.2 Méthode expérimentale

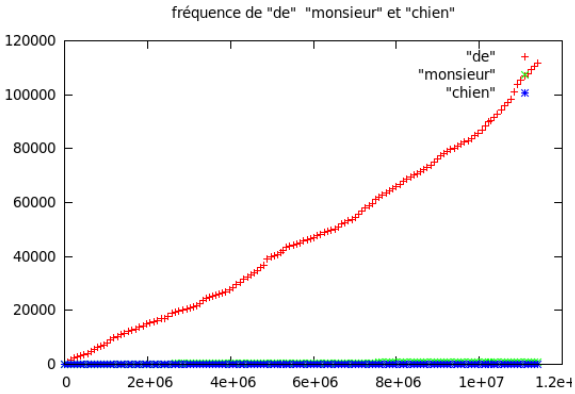
Une loi empirique comme celle de Zipf permet d'estimer si le contenu d'un texte est porteur de sens ou s'il relève d'une écriture aléatoire. Par contre, cette loi n'est pas adaptée à une analyse plus fine, car approximative et non discriminante pour les faibles valeurs de fréquence de mot (i.e classé en rang élevé dans la loi de Zipf). Nous proposons donc une nouvelle méthode d'analyse dynamique de corpus pour caractériser les probabilités d'occurrence des mots, et simultanément pour classer ces derniers en mots grammaticaux (mots " vides "), mots spécialisés de domaine ou mots courants :

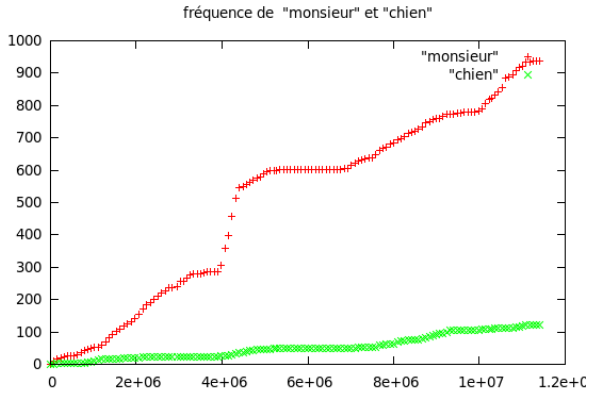
- Le corpus est analysé en flux continu. L'observation est réalisée périodiquement c'est-à-dire qu'on fige le comptage fréquentiel de tous les mots tous les k mots observés dans le corpus. Si, après avoir parcouru n premiers mots, on a décompté f_i occurrences d'un mot m_i alors $f_i \sim np_i$, où p_i est la probabilité d'apparition du mot m_i. D'après le précédent théorème, il suffit d'observer l'évolution de f_i en fonction de n quand n varie de 0 à taille maximale du corpus (que l'on considère comme très grand # ∞). En fonction de l'allure de la courbe $f_i(n)$ observée, on peut ensuite opter pour la loi décrivant le mieux la fréquence d'apparition du mot m_i.
- Si $f_i(x)$ tend vers une droite asymptote d'équation $y=c^{te}$ alors le théorème précédent s'applique. La distribution du mot m_i peut être alors modélisée par une loi de Poisson (et donc, pour les grandes valeurs de n, par une gaussienne à un seul paramètre λ). Dans une étude expérimentale, une courbe faiblement croissante, par exemple en log(x) peut aussi être acceptée comme une approximation acceptable de la droite asymptote $y=c^{te}$ (log-linéarité).

Expérimentalement, l'analyse de corpus «relativement» homogènes, comme le nôtre, montre que les fréquences des mots croissent plutôt linéairement. Nous retiendrons donc pour les grandes valeurs de n, un distribution gaussienne à 2 paramètres. De plus, l'analyse de la courbe d'évolution de chaque mot permet de classer ce dernier dans l'une des catégories précédemment évoquées. S'agissant d'un choix empirique des critères discriminants pour le classement, ce choix est justifié surtout par des observations dont l'exemple suivant est décrite en illustration.

3.3 Résultat expérimental et calcul de saillance

Les mots *de*, *monsieur*, *cheval* et *chien* ont été choisis pour représenter des classes de mots grammaticaux, de mots spécialisés et de mots d'emploi général. Leur courbe de fréquence cumulée est analysée sur notre corpus d'actualités. L'accroissement en fréquence du mot *de* est logiquement la plus rapide comme l'indique la figure ci-dessous. Comparativement, les courbes d'évolution des mots *chien* et *monsieur* paraissent plates. Ce n'est pas le cas comme l'indique la figure suivante lorsqu'on change le facteur d'échelle sur l'axe y. On constate aussi que la courbe de croissance du mot *de* est plus régulière autour de la droite qui la sous-tend tandis que les courbes de croissance des mots *monsieur* et *chien* sont plus dispersées.





On cherche donc à quantifier cette dispersion pour servir de critère de discrimination des mots à des fins de classement ou d'ordonnancement. La dispersion peut être traduite par la variance ou mieux, pour disposer d'une échelle de valeur uniformisée, par la forme normalisée qu'est le coefficient de variation.

Le calcul du coefficient de variation se fait comme suit : si f_1, f_2, ..., f_k désignent la suite de fréquences cumulées suivant le comptage décrit plus haut, et si $n_1,n_2,...,n_k$ désignent les nombres cumulés de mots parcourus pour décompter les f_i, alors la moyenne $\mu=\frac{f_k}{n_k}$ et la variance σ telle que $\sigma^2=(\sum_{i=1}^{k-1}\frac{f_{i+1}-f_i}{n_{i+1}-n_i}-\mu)^2$ permettent de calculer le coefficient de variation, égal à $\frac{\sigma}{\mu}$.

3.4 Utilisation du coefficient de variation

L'utilisation du coefficient de variation sur une échelle de valeur scalaire permet d'ordonner les mots (et les locutions une fois apprises) suivant un indice de notoriété. Intuitivement, ce ne sont pas les mots les plus fréquents qui présentent un intérêt mais plutôt ceux utilisés le plus régulièrement dans de nombreux contextes. En plus, en associant à chaque locution apprise sa catégorie grammaticale, et en se focalisant sur les catégories les plus porteuses d'information comme les groupes nominaux ou les patronymes, on arrive ainsi à extraire des éléments saillants mais évanescents comme *nuage de cendres*, *fuite de pétrole*...

3.5 Regroupement en catégories grossières

Le coefficient de variation permet d'estimer l'importance de chaque mot pris isolément par rapport au corpus. Expérimentalement, il permet une séparation effective des mots grammaticaux des autres mots. Mais pour regrouper les mots restants en catégories grossières, on a besoin de plus d'informations, et notamment de quantifier les interactions entre mots.

Pour pouvoir réutiliser les mêmes calculs expérimentaux que précédemment sur la fréquence des mots, et pour conserver une cohérence dans le formalisme de calcul, on remarquera qu'il existe un parallèle entre le coefficient de variation et la notion de tfxidf en recherche d'information (cette dernière correspond dans les modèles probabilistes à la probabilité d'avoir un document pertinent contenant un terme t). L'extension de cette mesure locale, liée à un document, vers une mesure globale sur le corpus se fait naturellement par la notion d'entropie $E(t)=\sum_{d\in D} -p_t\log(p_t)$. Plus un terme est uniformément distribué, plus sa valeur d'entropie est élevée. La notion d'entropie sur un terme isolé s'étend ensuite à celle sur des couples de termes t_1 et t_2 via la notion d'information mutuelle $I(t_1,t_2)=\sum_{d\in D} p(t_1,t_2)\log(\frac{p(t_1,t_2)}{p(t_1)p(t_2)})$. Plutôt que d'utiliser l'information mutuelle comme critère de regroupement des mots en catégories, on utilisera la notion de coefficient de corrélation linéaire entre couple de termes t_1 et t_2, qui est l'extension de la notion de coefficient de variation :

$\rho=\frac{\sigma_{t1,t2}}{\sigma_{t1}\sigma_{t2}}$ où $\sigma_{t1,t2}$ est la covariance de t_1 et t_2, et $\sigma_{t2}\sigma_{t2}$ leur variance respective .

Il s'agit *a posteriori* d'un calcul équivalent puisque $I(t_1,t_2)=-\frac{1}{2}\log(1-\rho^2)$.

La réalisation de cette partie est prévue pour la prochaine version du composant d'acquisition terminologique.

4 Analyse statistique de la compositionalité[2]

La compositionalité des mots est évaluée en linguistique par leur potentiel combinatoire. C'est un comptage fréquentiel, pour un mot donné, de l'appariement d'autres mots dans les constructions observées dans un corpus. Le potentiel combinatoire sert d'indicateur pour faciliter le travail d'analyse d'un lexicologue. Ce critère n'est cependant pas adapté à un apprentissage non supervisé, où l'analyse doit être réalisée automatiquement. Une notion plus appropriée concerne l'entropie, ce qui va être précisé ci-après.

Supposons que, dans un corpus, nous ayons observé 4 fois, le mot *bâton* dont deux avec le qualificatif *rouge*, une avec *bleu* et une avec *vert*. Si l'on souhaite ne garder qu'un seul indicateur qui résume la distribution, estimée à $(\frac{1}{2},\frac{1}{4},\frac{1}{4})$, la somme des probabilités présente peu d'intérêt comme indicateur. Par contre en étudiant la quantité d'information (Shannon, 1948) que chacun des précédents qualificatifs peut apporter au mot *bâton*, soit $(-\log(\frac{1}{2}),-\log(\frac{1}{4}),-\log(\frac{1}{4}))$ la moyenne attendue (espérance) est une bonne indication du degré de compositionalité du mot, soit $-\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{4}\log\left(\frac{1}{4}\right)-\frac{1}{4}\log\left(\frac{1}{4}\right)=1.5$ dans le cas d'une échelle logarithmique en base 2.

Cette valeur d'entropie indique la quantité d'information que peut recevoir en moyenne chaque mot. Si un mot m_1 a été observé n fois dans un corpus, son entropie a une valeur entre 0 et log(n). Une valeur nulle traduit l'existence d'un mot m_2 dont la probabilité d'observer en cooccurrence avec m_1, vaut 1. Le mot m_1 est dans ce cas non compositionnel puisque fortement lié à m_2. C'est le cas des mots comme *cochère*, *aujourd* , *lurette* ou *escampette*. A l'opposé, une valeur maximale de l'entropie, log(n), correspond à la distribution équiprobable c'est-à-dire à un fort degré de compositionalité. Normalement, c'est vers cette valeur maximale que tendent les mots grammaticaux.

4.1 Champ de compositionalité

Intuitivement, si un mot est employé dans son sens compositionnel, il est fort possible de trouver, dans le corpus, ce mot associé à d'autres mots à des degré divers. Par exemple *bâton* peut être associé à *rouge*, *vert*, *jaune*... et peut-être moins à *joyeux*, *espiègle*, *content*. Le champ de compositionalité d'un mot m correspond à une distribution probabiliste sur l'ensemble des mots m_i et traduit la probabilité d'observer m_i sachant qu'on a observé le mot m. Cette distribution peut être résumée par sa moyenne (entropie), sa variance et sa loi de distribution. Le champ ainsi défini permet d'introduire la notion d'intervalle de confiance et de déterminer de quelle manière une construction est jugée compositionnelle. Dans le cas d'une locution comme *retour de bâton*, il n'existe pas de forme altérée ou modifiée ne comportant qu'une partie de mots de ce groupe. Si la locution est souvent employée dans le corpus, la fréquence d'association est plus élevée que le cas des constructions compositionnelles, ce qui doit permettre à une analyse statistique de conclure que l'un des mots *retour* ou *bâton* n'appartient pas au champ compositionnel de l'autre mot.

4.2 Méthode expérimentale

Pour chaque mot m_0, l'entropie et la variance sont déduites expérimentalement à partir d'un comptage fréquentiel de cooccurrences :

- Pour chaque mot m_0, on procède au comptage de cooccurrence $f_{0,i}$ (resp. $f_{i,0}$) des mots m_i contigus au mot m_0 à droite (resp. à gauche). Le positionnement gauche/droite reflète la nature séquentielle du corpus de texte.
- Pour évaluer le degré de compositionalité, le comptage ne devrait porter que sur les mots m_i ayant un sens compositionnel avec le mot m_0 et exclure les mots m_i lorsque m_0m_i constitue une locution. Au stade de l'apprentissage, on ne dispose pas d'une telle information. L'hypothèse est que les mots m_i constituant une locution sont en plus faible nombre que les mots m_i portant un sens compositionnel. Cela justifie l'approximation dans l'estimation de la moyenne et de la variance.
- Pour tenir compte de la masse absente (due au manque d'exhaustivité de tout corpus), on procède à un lissage de Laplace. La valeur de lissage est plus petite que 1, en raison des faibles fréquences de cooccurrence.

[2] Les notions de compositionalité, de champ de compositionalité... sont revues ici dans une logique calculatoire

- La probabilité $p_{0,i}$ d'observer le mot m_i est estimée par $p_{0,i}=\frac{f_{0,i}}{\sum_j f_{0,j}}$. L'entropie μ_0 et la variance σ_0 sont estimés par $\mu_0=-\sum_j p_{0,j}\log p_{0,j}$ et $\sigma_0^2=\sum_j p_{0,j}(\mu_0-\log p_{0,j})^2$.

4.3 Modélisation de la loi de compositionalité

Il reste à déterminer la loi de distribution des log $p_{0,i}$ pour pouvoir fixer l'intervalle de confiance à l'intérieur duquel une association de mots est considérée comme compositionnelle.

- Pour chaque mot m_0, la cooccurrence d'un mot m_i peut être considérée comme une épreuve de Bernoulli et la fréquence de cooccurrence comme une v.a.r X_i de loi binomiale de paramètre p_i.
- Pour les grandes valeurs de fréquence, la loi de X_i peut être approximée par une loi gaussienne. Nous nous intéressons pour la suite à la v.a.r $\frac{X_i}{n_0}$ où n_0 est le nombre total de cooccurrences observées pour le mot m_0. $\frac{X_i}{n_0}$ suit également une loi gaussienne que nous désignerons par X.

Pour la suite, μ_0 peut être considérée comme un résultat d'observation d'une v.a.r $Y=\sum_i \frac{X_i}{n_0}\log(\frac{X_i}{n_0})$. On est donc amené à étudier en premier la loi de X log(X) connaissant la loi de X.

4.4 Approximation de Y par une gaussienne.

X étant une distribution connue, on cherche, pour ce faire, à déterminer la fonction de distribution g de la v.a.r Y=X log(X) à partir de la fonction de distribution f de X.

La démarche classique consiste à évaluer à partir de F, fonction de répartition de X, la fonction de répartition G de Y alors : $G(y)=P(Y<y=\varphi(x))$ avec $\varphi(x)=-x\log(x)$.

La fonction φ n'est pas bijective. Elle est définie, s'agissant de valeurs de probabilité, sur l'intervalle $[0,1]$. Elle est croissante sur $[0,\frac{1}{e}]$ et décroissante sur $[\frac{1}{e},1]$.
La fonction inverse φ^{-1} est déterminée graphiquement à partir de φ par la symétrie axiale par rapport à la droite d'équation y=x. φ^{-1} est bivaluée et elle est composée d'une branche strictement croissante $\varphi_1^{-1}:[0,\frac{1}{e}]\rightarrow[0,\frac{1}{e}]$ et d'une branche strictement décroissante $\varphi_0^{-1}:[0,\frac{1}{e}]\rightarrow[\frac{1}{e},1]$.
Plus précisément, si W_0 et W_{-1} sont les branches définies sur $[-\frac{1}{e},1]$ de la fonction W de Lambert[3], partie réelle, alors $\varphi_0^{-1}(y)=\frac{-y}{W_0(y)}$ et $\varphi_1^{-1}(y)=\frac{-y}{W_1(y)}$. Par suite:

$P(Y<y)=P(X<\varphi_1^{-1}(x))+1-P(X>\varphi_0^{-1}(x))$ ce qui peut s'écrire : $G(y)=F(\frac{-y}{W_{-1}(y)})+1-F(\frac{-y}{W_0(y)})$. La dérivée de W étant $W'(y)=\frac{W(x)}{x(1+W(x))}$, les dérivées de $\frac{-y}{W_{-1}(y)}$ et de $\frac{-y}{W_0(y)}$ sont respectivement $\frac{-1}{1+W_{-1}(y)}$ et $\frac{-1}{1+W_0(y)}$ d'où la fonction de distribution : $g(y)=\frac{-f(\frac{-y}{W_{-1}(y)})}{1+W_{-1}(y)}+\frac{f(\frac{-y}{W_0(y)})}{1+W_0(y)}$ où g et f sont les dérivées respectives de G et F.

La fonction de Lambert est difficile à mettre en œuvre dans un calcul numérique du fait des phénomènes d'oscillation lorsqu'on doit utiliser son développement en série limitée. Nous nous contenterons donc de rechercher l'allure générale de la courbe g(y) afin de l'approximer par une fonction plus simple.

[3] la fonction de Lambert peut être visualisée ici : http://math.asu.edu/~kawski/MAPLE/274/images/Lambert8.gif

Domaine de variation de g

La fonction $g(y)=\frac{-f(\frac{-y}{W_{-1}(y)})}{1+W_{-1}(y)}+\frac{f(\frac{-y}{W_0(y)})}{1+W_0(y)}$ est définie sur [0,1/e] et de domaine de variation [0,1].

Pour $y \to 0$, $\frac{-1}{1+W_{-1}(y)} \to 0$ et $\frac{-1}{1+W_0(y)} \to 1$

Pour $y \to \frac{1}{e}$, $\frac{-1}{1+W_{-1}(y)} \to \infty$ et $\frac{-1}{1W_0(y)} \to \infty$

Si maintenant f est une partie gaussienne définie sur [0,1], f est associée à $W_{-1}(y)$ sur [0,1/e] et à $W_0(y)$ sur [1/e,1], 3 cas de figures se présentent suivant que l'espérance μ et la variance σ de la fonction f conduisent à un recouvrement important de la valeur critique 1/e par la gaussienne définie par f.

- Pour $\mu \ll 1/e$, c'est la composante $\frac{-f(\frac{-y}{W_{-1}(y)})}{1+W_{-1}(y)}$ dans g qui est prédominante. Par suite g peut être approximée par une gaussienne avec une asymétrie (skew négatif) d'autant moins marquée que μ est proche de 0.
- De manière similaire pour μ proche de 1, c'est la composantedans g $\frac{f(\frac{-y}{W_0(y)})}{1+W_0(y)}$ qui est prédominante. Et par suite g peut être approximée par une gaussienne avec une asymétrie (skew positif) d'autant moins marquée que μ est proche de 1.
- Dans le cas d'un recouvrement conséquent de la valeur critique 1/e par la gaussienne, l'allure de la distribution g nécessite une analyse approfondie, autour de 1/e, du comportement joint de $f(\frac{-y}{W_{-1}(y)})$ modulé par $1+W_{-1}(y)$ d'une part, et de $f(\frac{-y}{W_0(y)})$ modulé par $1+W_0(y)$ d'autre part. Ce cas ne sera pas traité ici.

En pratique, nous ne nous intéresserons qu'au premier cas, où $\mu << 1/e$. En effet, la taille d'un vocabulaire type est de 50000 à 300000 mots (sans distinction des catégories grammaticales). Ce qui fait, dans nos estimations de μ à partir d'un comptage fréquentiel, et en effectuant un lissage de Laplace pour prendre en compte la masse absente, que la valeur de μ est très éloignée de 1/e et plutôt proche de 0. La représentation de la distribution g par une gaussienne est dans ce cas justifiée.

Si, maintenant, X_1 et X_2 sont 2 v.a.r de loi f_1 et f_2, alors la loi de X_1+X_2 est le produit de convolution f_1*f_2. Et dans le cas où X_1 et X_2 sont des gaussiennes, X_1+X_2 est aussi une gaussienne. En fonction des calculs estimatifs précédents et dans les conditions de nos expérimentations, nous admettrons que $Y=\sum_i \frac{X_i}{n_0}\log(\frac{X_i}{n_0})$ peut être approximée par une loi gaussienne.

4.5 Mise en œuvre de l'identification de non-compositionalité

Le comptage fréquentiel décrit dans §4.2 permet d'associer à chaque mot, pris individuellement, des caractéristiques de compositionalité à droite (resp. à gauche) via la moyenne et la variance. L'hypothèse d'une distribution gaussienne permet ensuite de définir un intervalle de confiance fixé expérimentalement à 95% (ce qui correspond à une valeur de 1.96 d'écart pour une gaussienne).

Pour tout mot m_1 de moyenne «à droite» $\mu_{d,1}$ et de variance «à droite» $\sigma_{d,1}$, si m_1 est suivi de m_2, de moyenne «à gauche» $\mu_{g,2}$ et de variance «à gauche» $\sigma_{g,2}$, m_1m_2 est non compositionnel:

- si $-\log(p_{dg,12}) < \mu_{d,1} - 1.96x\,\sigma_{d,1}$ où $p_{dg,12}$ est la probabilité d'avoir le mot m_2 qui suit le mot m_1
- ou si $-\log(p_{gd,21}) < \mu_{g,2} - 1.96x\sigma_{g,2}$ où $p_{gd,21}$ est la probabilité d'avoir le mot m_1 qui précède le mot m_2

5 Couplage du modèle linguistique

Le composant linguistique dispose au départ :

- d'un lexique du français comportant 300 000 formes fléchies, décrites par la partie du discours et des traits d'accord
- de règles de grammaires de chunking (Abney, 1994) de type hors contexte, décrites sous forme normale de Chomsky et regroupées par paquets homogènes
- de méta-règles régissant les paquets de règles afin de rendre, autant que possible, l'analyse déterministe.

De plus, la profondeur d'analyse est limitée pour couvrir des syntagmes de moins de 6 mots, ce qui est suffisant dans nos applications. Cette hypothèse permet de traduire les règles initiales en règles de grammaires régulières au sein de chaque paquet de règles.

Une première analyse lexicale du corpus permet de recenser le vocabulaire utilisé et de compléter le référentiel lexical initial par les nouveaux mots simples inconnus. L'ajout de ces mots inconnus dans le référentiel lexical est réalisé seulement après seuillage suivant leur fréquence d'occurrence et leur coefficient de variation.

5.1 Analyse lexicale et syntaxique du corpus

S'agissant de grammaires de chunking, l'absence du non-terminal initial S impose une analyse «bottom-up». Il s'agit donc d'une analyse LR classique (Aho et al., 1977) avec une utilisation particulière du chart parsing.

En effet, plutôt que de créer un espace de chart pour l'analyse de chaque phrase du corpus, on construit successivement des niveaux de chart couvrant tout le corpus et de la manière suivante :

- on dispose d'un référentiel lexical de mots simples et de locutions, et d'un référentiel des syntagmes en cours de construction
- le référentiel des syntagmes est vide au départ (éventuellement celui des locutions aussi)
- le référentiel lexical et le référentiel des syntagmes sont utilisés pour indexer tout le corpus
- le résultat de chaque indexation correspond alors à un niveau du chart
- une analyse du coefficient de variation des syntagmes du référentiel permet d'éliminer les éléments les moins pertinents
- une analyse de la compositionalité des syntagmes figurant dans le référentiel des syntagmes permet d'identifier les locutions et de les reverser dans le référentiel des locutions
- on applique ensuite un nouveau paquet de règles de grammaires pour identifier de nouveaux syntagmes et pour les reverser dans le référentiel des syntagmes

Le processus se termine après épuisement des paquets de règles.

5.2 Mise en œuvre du système

Les résultats qui suivent sont issus du corpus d'actualités décrit précédemment dans §2. Les données, en constante évolution, correspondent aux actualités de janvier 2011. Les listes ci-après correspondent à des extraits de patronymes et de groupes nominaux classés par ordre de pertinence décroissante. Un référentiel terminologique unique est dans un premier temps appris sur le corpus global d'actualités puis «projeté» sur des plus petits corpus thématisés, par analyse du coefficient de variation *intra* corpus.

Patronymes culturel	GN culturel	Patronymes sport	GN sport	Patronymes international	GN international
-nicolas sarkozy -johnny hallyday -frédéric mitterrand -conrad murray -dany boon -ben ali -john barry -brice taton -robert de niro - luc chatel	-golden globes -homicide involontaire -premier ministre -discours d un roi -grand palais -los angeles -première fois -poivre d arvor -sol majeur	-andy murray -caroline wozniacki -paris sg -jean pierre dick -claude onesta -kim clijsters -justine henin -wilfried tsonga -cyril despres -stanislas wawrinka	-autres sports -quarts de finale -championnats étrangers -tête de série -finale de la coupe -championnat du monde -fin de la saison -milieu de terrain	-ben ali -laurent gbagbo -nicolas sarkozy -sidi bouzid -zine ben -saad hariri -vincent delory -jean claude duvalier -silvio berlusconi	-premier ministre -affaires étrangères -service français -ancien président -départ du président -président déchu -président tunisien -conférence de presse -forces de l ordre

-xavier beauvois -beverly hills -marc olivier fogiel -sofia coppola -justin bieber -quentin tarantino -laurent gerra -caroline lachowsky -claude monet -françois fillon -ernest hemingway -alexandre jardin -jean dutourd	-biographie d hemingway -priorité santé -haute couture -bande dessinée -mise en scène -accusé de plagiat -tête de bois -meilleur film -jeu vidéo -nouvelles technologies -télé réalité -pluies diluviennes -premier album -bande dessinée d angoulême	-josé mourinho -michel desjoyeaux -stéphane sessegnon -paris fc -françois gabart -jean tigana -saint etienne -loïck peyron -carlos sainz -dimitri payet -brian joubert -tomas berdych -lionel messi	-ballon d or -journal du mercato -finale du tournoi -champion du monde -français jean -coupe de la ligue -ski alpin -match en retard -nuit des français -rumeurs du mercato -tenant du titre -quart de finale -nuit dernière -première fois -conférence de presse	-benoît xvi -françois fillon -mohamed elbaradei -alain juppé -nelson mandela -gilles trequesser -mohamed ghannouchi -antoine de léocour -eric zemmour -jean stéphane -johan vande -tarek amara -henri pierre -eric faye	-régime du président -ministre des affaires -nouveau gouvernement -journaliste de l afp -droits de l homme -jeunes français -président américain -ministère de l intérieur -démission du gouvernement -président sortant -union européenne -ministre de la défense -communauté internationale -français enlevés

6 Conclusion

L'approche que nous venons de décrire confirme qu'il est possible de concevoir un système d'acquisition de terminologie performant en temps d'exécution et aussi de très bonne qualité. Le taux de précision[4] obtenu est de l'ordre de 90% (Lassalle et al., 2011). Les principales raisons de ces performances sont liées à :

- une architecture de chart parsing couvrant tout le corpus, évitant ainsi des redondances d'analyse des mêmes syntagmes
- le regroupement des syntagmes analysées dans un même référentiel, permettant ainsi un couplage avec l'analyse statistique tout en minimisant le biais
- une spécialisation des analyses statistiques entre la détection des locutions et le classement de ces dernières en fonction du corpus applicatif

7 Annexe :

Extrait de la grammaire de chunking permettant d'identifier les patronymes :

#Cat prenoms.prenoms

- (CatLoc1 prenoms.prenoms) →(CatMot prenoms) (CatMot prenoms)
- (CatLoc1 prenoms.prenoms) →(CatMot particule.prefixe) (CatMot prenoms)

#Cat PRENOMS.particule

- (CatLoc1 PRENOMS.particule) →(CatMot prenoms) (CatMot particule)
- (CatLoc1 PRENOMS.particule) →(CatMot prenoms) (CatLoc1 particule.particule)
- (CatLoc1 PRENOMS.particule) →(CatLoc1 prenoms.prenoms) (CatMot particule)
- (CatLoc1 PRENOMS.particule) →(CatLoc1 prenoms.prenoms) (CatLoc1 particule.particule)

#syntagme PATRO
avec détection de non-compositionalité

- (CatLoc1 PATRO) →(CatLoc1 PRENOMS.particule) (CatMot patronyme) + (SeuilleOr $LOCBIN1)
- (CatLoc1 PATRO) →(CatLoc1 PRENOMS.particule) (CatMot prenoms) + (SeuilleOr $LOCBIN1)
- (CatLoc1 PATRO) →(CatLoc1 PRENOMS.particule) (CatMot v.stat) + (SeuilleOr $LOCBIN1)
- (CatLoc1 PATRO) →(CatLoc1 prenoms.prenoms) (CatMot patronyme) + (SeuilleOr $LOCBIN1)

[4] Le taux de rappel n'est pas pertinent pour un modèle d'apprentissage statistique. En effet, un nombre minimal d'occurrences (environ 4) d'une même locution est nécessaire pour que cette dernière puisse être identifiée, ce qui exclut des locutions dont la fréquence d'apparition est trop faible. Enfin, l'estimation de ce taux nécessite un recensement manuel des locutions dans le corpus de test, ce pour un coût en général prohibitif. Une solution (que nous n'avons pas mise en œuvre) consisterait à échantillonner le corpus pour estimer le nombre moyen de locutions observées tous les n mots analysés et de le comparer avec le nombre total des locutions extraites divisé par la taille (en nombre de mots) du corpus d'apprentissage.

- (CatLoc1 PATRO) →(CatLoc1 prenoms.prenoms) (CatMot v.stat) + (SeuilleOr $LOCBIN1)

Références

ABNEY S.T.,(1994). PARSING BY CHUNKS. *BELL COMMUNICATION RESEARCH*.

AHO A.,SETHI R., ULLMAN J.D.(1977). Compilers: Principles, Techniques, and Tools. *Dragon Book*.

BRILL E.,(1992). A Simple Rule Based Part of Speech Tagger. *ACL*.

CHURCH K., HANKS P.,(1996). WORD ASSOCIATION NORMS, MUTUAL INFORMATION, AND LEXICOGRAPHY. *COMPUTATIONAL LINGUISTICS.* 16, 22-29.

CORLESS ET AL.,(1996). On the Lambert W function. *Adv. Computational Maths.* 5, 329-359.

DAILLE B.,(1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *MIT Press.*, 49-66.

DUNNING T.D.,(1993). Accurate Methods for the Statistics. *Computational Linguistics. 19(1)*, 61-74.

FRANTZI K.T., ANANIADOU S., TSUJII J.,(1998). The C-value/NC-value Method of Automatic Recognition of Multi-word Terms. *ECDL'98*, 585-604.

HEINECKE J., SMITS G., CHARDENON C., GUIMIER DE NEEF E.,MAILLEBUAU E., BOUALEM M., (2008). TILT : plate-forme pour le traitement des langues naturelles. *TAL Vol. 49*.

KIT C.,(2002). Corpus Tools for Retrieving and Deriving Termhood Evidence. *The 5th East Asia Forum of Terminology*, 69-80.

LASSALLE E., CASIMIR P.K., GUIMIER DE NEEF E.,(2011). Evaluation des outils d'extraction terminologique Quezao et Acabit. *EGC 2011*, 131, 136.

NAMER F.,(2000). Flemm : Un analyseur flexionnel de français à base de règles. *Traitement Automatique des Langues pour la Recherche d'Information. Hermes*, 523-547.

NAZARENKO A., ZARGAYOUNA H., HAMON O., VAN PUYMBROUCK J.,(2009). Evaluation des outils terminologiques : enjeux, difficultés et propositions. *TA Vol. 50*, 257-281.

PAPOULIS A.,(2002). Probability, Random Variables ans Stochastic Processes. *Mac Graw Hill*.

SAPORTA G., (2006). Probabilité, analyse des données et statistique. *Ed. Technip*.

SHANNON C.E.,(1948). A Mathematical Theory of Communication. *The Bell System Technical Journal. 27*, 623-656.

SMADJA F.,(1993). XTRACT : An Overview. *Computer and the Humanities Kluwer Academic Publishers*.

TSURUOKA Y.,(2005). Chunk Parsing Revisited. *9th IWPT*.

VU T., AW A.T., ZHANG M.,(2008). Term Extraction Through Unithood And Termhood Unification. *IJNLP*.