Evaluation des outils d'extraction terminologique Quezao et Acabit

Edmond Lassalle*, Prem Kumar Casimir**
Emilie Guimier de Neef***

Orange Labs, 2 av. Pierre Marzin 22300 Lannion France
{edmond.lassalle, premkumar.casimir, emilie.guimierdeneef}@orange-ftgroup.com

Résumé. L'article décrit l'évaluation de deux outils d'extraction terminologique Acabit et Quezao. Si Acabit est plus connu car librement disponible, Quezao est issu des travaux d'Orange Labs sur la recherche d'informations. Après une comparaison sur les approches théoriques des deux systèmes, une évaluation concrète va porter sur un corpus d'actualité (2424Actu) pour l'aspect qualitatif et sur un corpus de presse pour l'aspect quantitatif (dont la tenue en charge).

1 Introduction

L'évaluation d'outils d'extraction terminologique (monolingue), en particulier pour le français, est à ce jour peu répandu [Nazarenko et al.]. La spécificité des langues, l'absence d'une fédération internationale sont en effet des freins à une campagne d'évaluation importante du même niveau que TREC [http://trec.nist.gov/].

Les rares campagnes d'évaluation comme ARC A3 [El Hadi et al.] ou CESART [Timimi] mettent en évidence la difficulté de comparer différents systèmes du fait de la diversité des fonctionnalités proposées. Des 9 systèmes participant à la campagne CESART (IDE/XTS, Lexter, WorldTrek, Termic, Termos, SeekJava, SynoTerm et Terminae, TermWatch), moins d'un système sur deux pouvait partager le même protocole d'évaluation. Enfin, le coût élevé de réalisation des corpus d'évaluation, en limitant leur nombre et le domaine traité, ne permet pas une estimation correcte de chaque outil en conditions réelles d'utilisation.

Le cadre de nos travaux sur Quezao et la présente évaluation sont dictés par des besoins internes à Orange. Il s'agit d'applications de Recherche d'Informations (RI), couvrant des domaines variées, de la documentation technique interne au contenu grand public (moteur de recherche généraliste ou dédié). L'évaluation doit porter sur des corpus d'applications réelles d'Orange. Elle a dû être menée en interne car les données propriétaires ne pouvaient être mises à disposition dans une campagne d'évaluation associant des partenaires externes. L'intervention humaine dans l'évaluation a donc été réduite, et consiste seulement à valider les termes produits en sortie de chaque système. L'absence de jeu de validation issu d'un corpus de test est cependant acceptable dans une évaluation de type «boîte transparente» où nous cherchons aussi à connaître le fonctionnement interne de chaque système.

2 Choix d'Acabit pour une comparaison avec Quezao

Les systèmes d'extraction terminologique (STE) récents sont basés sur la stratégie C/NC-

value [Frantzi et al] qui consiste à utiliser :

- une analyse linguistique du corpus pour identifier les constructions syntaxiques valides pouvant correspondre à des locutions (candidats termes)
- une analyse statistique des candidats termes pour les classer suivant un degré de vraisemblance (C-value) d'être une locution et un degré de spécificité au domaine applicatif visé (NC-value).

Cette stratégie est revue et généralisée dans [Kit] par les notions d'unithood (C-Value) et de termhood (NC-value). Les mesures de classement d'unithood peuvent être l'information mutuelle, la log-perplexité, la log-vraisemblance, etc. tandis qu'un modèle markovien (n-gram) est proposé pour le calcul (contextuel) de termhood.

Cependant, les STE ne mettent pas l'accent sur le traitement linguistique, ni sur son impact non négligeable dans les performances et dans la qualité des systèmes. Le faible nombre d'outils linguistiques disponibles pour le français (LEXTER, INTEX ou SYNTEX), les campagnes d'évaluation suivant une approche «boîte noire» ne permettent pas, pour ces raisons, de mesurer séparément l'apport du traitement linguistique et celui du traitement statistique dans les systèmes.

De plus, de par le large éventail d'applications chez Orange, la réduction du coût d'adaptation à chaque nouveau déploiement est un impératif dans la conception de Quezao. Les connaissances renseignées manuellement (comme les données linguistiques) doivent être réduites à leur strict minimum. L'évaluation cherchera donc à valider également l'hypothèse qu'un faible apport linguistique peut être compensé par une analyse statistique poussée autre que celle du C/NC-value qui prédomine. Ceci ne peut être fait que par une connaissance fine du fonctionnement du STE que l'on veut évaluer.

Le choix d'Acabit comme système à comparer à Quezao est donc motivé par sa disponibilité, son architecture ouverte permettant d'apparier différents outils d'analyse linguistique, son approche conforme à la philosophie C/NC-value. De plus, bien que les modèles probabilistes diffèrent, l'hypothèse initiale est partagée en considérant que les occurences des mots dans un corpus suivent une loi binomiale, ce au vu d'une observation expérimentale [Dunning].

3 Principe de fonctionnement des 2 systèmes

L'extraction terminologique des 2 systèmes utilise l'analyse linguistique pour identifier les constructions syntaxiques et une analyse statistique pour classer (Acabit) ou seuiller les candidats termes (Quezao). Le classement correspond à la stratégie C/NC-value, le seuillage automatique n'est pas à notre connaissance utilisé à ce jour.

3.1 Modèle Acabit

Acabit produit une liste de candidats termes classés suivant la log-vraisemblance :

- chaque mot est associé au nombre de ses occurrences dans le corpus d'apprentissage
- ce nombre est décrit par une variable aléatoire suivant une loi binomiale de paramètre p
- deux mots contigus constituent une locution si statistiquement, ils «apparaissent fréquemment ensemble».
- sous contrainte de contiguïté, cela signifie que les variables aléatoires associées sont dépendantes.
- sous contrainte de contiguïté et d'une modélisation par une loi binomiale, cela signi-

fie que les paramètres de loi, soient p_1 et p_2, sont identiques

- l'estimation de p_1 et p_2 est réalisée par le comptage fréquentiel suivant :

$$-2\log\lambda = 2\left[\log L(p_{1,}k_{1,}n_1)+\log L(p_{2,}k_{2,}n_2)-\log L(p,k_{1,}n_1)-\log L(p,k_{2,}n_2)\right]$$

$$où\ \log L(p,n,k)=k\log p+(n-1)\log(1-p)\ et\ p_1=\frac{k_1}{n_{1,}}\ p_2=\frac{k_2}{n_2}\ et\ p=\frac{k_1+k_2}{n_1+n_2}$$

$$et\ k_1=f(AB), k_2=f(\sim AB), n_1=f(AB)+f(A\sim B)\ et\ n_2=f(\sim AB)+f(\sim A\sim B)$$

$f(AB), f(\sim AB), f(A\sim B)$ *et* $f(\sim A\sim B)$ *désignant respectivement les*:

. présence conjointe de A et B

. présence conjointe d ' un mot différent de A avec le mot B

. présence conjointe de A avec un mot différent de B

. présence conjointe de 2 mots différents de A et de B respectivement

Plus la valeur de -log λ est faible, plus les 2 v.a.r A et B sont fortement dépendantes. -log λ est enfin fortement dépendante des caractéristiques du corpus selon que ce dernier est plus ou moins fourni en terminologie. Pour un corpus chargé, la valeur de seuillage est plus élevée que pour pour un corpus comportant peu de terminologie.

Par ailleurs, Acabit est un système ouvert en terme de modularité, en facilitant le traitement en amont par n'importe quel module linguistique moyennant peu de contraintes. Mais cela se traduit en contrepartie par l'impossibilité par le module statistique de catégoriser grammaticalement les locutions extraites, faute d'informations linguistiques adéquates (relatives à la structure syntagmatique).

Enfin, si la log vraisemblance permet d'ordonner les termes obtenus, il existe dans ce classement une importante plage intermédiaire où des termes compositionnels sont mélangés indistinctement à des locutions. Il est donc difficile de fixer un seuillage sans nuire à l'exactitude ou à la complétude.

3.2 Modèle Quezao

Le principe de fonctionnement de Quezao est basé sur la modélisation de la compositionnalité des mots. Lorsqu'un mot est utilisé dans le sens compositionnel, le sens de son emploi est affiné par le mot qui suit ou le précède dans le texte. Cela peut être évalué en terme de quantité d'information. La probabilité est estimée expérimentalement par un comptage fréquentiel :

- La linéarité du texte va se traduire par deux quantités d'informations reçues à la gauche et à la droite du mot respectivement.
- Si A est un mot suivi d'un mot B, la quantité d'information à droite de A est -log P(B|A précède B) que l'on note par -log P_A(B), P(.|A précède B) étant en effet une probabilité.
- Pour chaque mot A, on va estimer la distribution -log(P_A(B)) pour tous les mots B du vocabulaire. Le calcul suppose que AB est employé dans le sens compositionnel. Aucune analyse ne permettant à ce stade de savoir si tous les AB sont utilisés dans leur sens compositionnel, l'hypothèse est que les séquences AB correspondant à des locutions sont en plus faible nombre que les séquences utilisées dans leur sens compositionnel. Comme les locutions ne devraient pas en principe intervenir dans ce calcul de compositionnalité, l'hypothèse d'un plus faible nombre relatif de locutions suppose que l'erreur induite dans le calcul est faible.

- Cette distribution va permettre de déterminer l'espérance (l'entropie gauche et droite) de chaque mot A du vocabulaire.
- On estimera également la variance de la distribution. Enfin, en supposant que la probabilité pour A de recevoir une quantité d'information suit une loi de Gauss, la distribution de probabilité en question est alors entièrement connue, ce qui permet de fixer un intervalle de confiance de la valeur d'entropie pour A.
- Pour chaque mot A fixé, son modèle de distribution de quantité d'information étant connu avec son intervalle de confiance, le comptage des occurrences d'un mot B succédant A va fournir l'estimation de l'apport d'une certaine quantité d'information par B. Si cette quantité d'information est en dehors de l'intervalle de confiance fixant l'entropie de A, on considère que la quantité d'information apportée par B est nulle et que AB constitue dans ce cas une locution.
- La notion d'intervalle de confiance va permettre un seuillage individualisé pour chaque mot. Il n'y a pas de classement des termes suivant un ordre de vraisemblance comme dans Acabit puisque tout terme proposé par Quezao est supposé être une locution.

Le modèle linguistique dans Quezao est imposé pour permettre la catégorisation grammaticale des termes produits. En effet, les modules linguistique et statistique interagissent dans un fonctionnement en parallèle pour valider et classer les constructions syntaxiques intermédiaires et finales. Pour faciliter l'adaptation à d'autres langues, il s'agit d'un modèle linguistique minimal (un lexique typé uniquement par les parties du discours - nom, verbe, adjectif.. - et, pour décrire les locutions, une grammaire de chunk [Abney] dans une stratégie d'analyse LR rendue déterministe par des méta-règles de précédence). La simpllicité du modèle permet une complexité d'analyse quasi-linéaire comparée à une complexité habituelle entre $o(n^2) et o(n^3)$ où n est la longueur moyenne des phrases. Le tableau suivant confirme les bonnes performances en charge de cette approche.

	Quezao	Acabit
Corpus de 10 Mo	45sec	7min40
Corpus applicatif : 2424Actu–95 Mo	6 min	152 min
Corpus applicatif :Presse et Média 2008 - 758 Mo	137 min	overflow

3.3 Classement par vraisemblance, classement par notoriété

A ce stade, Quezao ne propose pas comme Acabit un classement des termes. L'ordre de pertinence qui va être établi consiste, dans une logique de recherche d'information, à utiliser et à étendre la notion de tfxidf. Celle-ci 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. L'extension de la notion de terme isolé vers des couples de termes t_1 et t_2 se fait ensuite via la notion d'information mutuelle

$$I(t_1,t_2) = \sum_{d \in D} p(t_1,t_2) \log\left(\frac{p(t_1,t_2)}{p(t_1)\,p(t_2)}\right) \quad .$$

C'est cette méthode qui est utilisée mais sous une forme différente, basée sur la régularité d'emploi du mot. La régularité n'est pas forcément la fréquence d'emploi et elle est déterminée par le coefficient de variation calculé pour chaque mot donné sur un corpus donné.

Le corpus étant supposé 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. Le calcul du coefficient de variation se fait comme suit :

Si $f_1, f_2, \ldots, f_k$ *désignent la suite des fréquences cumulées et si* $f^l_1, f^l_2, \ldots, f^l_k$ *désignent les fréquences locales dans chaque intervalle ,* $f^l_1 = f_1, f^l_i = f_i - 1 \ldots, f^l_k = f_k - 1$ *suivant la méthode décrite , et si* $n_1, n_2, \ldots, n_k$ *désignent le nombre de mots parcourus pour décompter les* f_i *, alors la moyenne* $\mu = \frac{f^l_k}{n_k}$ *et la variance* σ *est telle que* $\sigma^2 = \sum_{i=1}^{k} \left(\frac{f^l_i}{n_i} - \mu \right)^{2.}$ *Le coefficient de variation est alors égal à* $\frac{\sigma}{\mu}$

L'extension vers un couple de termes t_1 et t_2 correspond au coefficient de corrélation linéaire $\rho = \frac{\sigma_{t1,t2}}{\sigma_{t1}\,\sigma_{t2}}$ *où* $\sigma_{t1,t2}$ *est la covariance de* t_1 *et* t_2, *et* $\sigma_{t2}\,\sigma_{t2}$ *leur variance respective*

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

Un terme est considéré comme pertinent si son coefficient de variation est faible. En associant une catégorisation grammaticale et en utilisant un réseau sémantique construit automatiquement, Quezao propose des associations comme *Nicolas Sarkozy avec réforme des retraites* ou *discours de Grenoble* ou *sécurité* au fil de l'actualité.

4 Evaluation des deux systèmes

Le nombre de termes à valider manuellement étant élevé (plus de 30000 pour chaque système), on se contentera d'analyser les 1000 premiers termes proposés par chaque système.

Taux de précision : il est calculé pour chaque n fois 100 premiers termes, pour n variant de 1 à 10. C'est le ratio du nombre de termes valides (*i.e* considérés comme locution) sur le nombre total de n fois 100 termes.

Taux de recouvrement : on ne cherchera pas à déterminer le taux de rappel face à une impossibilité de dénombrer la terminologie issue d'un corpus. Une solution consisterait à extraire du corpus, par tirage aléatoire, un sous ensemble représentatif de documents et à estimer le taux de rappel sur la base de cet échantillon. Nous ne retiendrons pas pour le moment cette méthode plus longue à mettre en œuvre. De plus, la méthode risque de fausser l'évaluation du classement préférentiel des 2 systèmes.

Le taux de recouvrement que nous utilisons à la place permet de comparer directement les systèmes plutôt que de les placer sur une échelle de valeur absolue.

Pour chaque n fixé, on se contentera de dénombrer n_Q et n_A termes valides issus de Quezao et Acabit ainsi que $n_{A\cup Q}$ termes valides issus de Quezao ou d'Acabit.

Le taux de recouvrement de Quezao et Acabit est respectivement $n_Q/n_{A\cup Q}$ et $n_A/n_{A\cup Q}$. Ce taux indique sur l'ensemble des termes produits par les 2 systèmes, la part relative de couverture de chaque système.

La chaîne de traitement pour Acabit comprend pour la partie linguistique l'étiqueteur de Brill [Brill.93] et l'outil Flemm [Namer]. Dans une deuxième partie de l'évaluation, Acabit est couplé avec l'analyseur Tilt d'Orange [Heinecke].

Le taux de précision (resp. recouvrement) de Quezao est de plus de 90% (resp. 70%) pour les 1000 premiers termes proposés et pour Acabit d'un peu moins de 50% (resp. 50%) avec Brill+Flemm et de 80% (resp. 60%) avec Tilt.

5 Conclusion

L'évaluation a mis en évidence l'impact important du modèle linguistique dans les STE se conformant à la stratégie C/NC-value. La mesure de log-vraisemblance d'Acabit, même datée, reste compétitive dans l'évaluation. Enfin, avec un modèle linguistique simple dans Quezao, il est possible d'avoir une bonne qualité des résultats en optant pour une stratégie de seuillage et non de classement préférentiel. La tenue en charge de Quezao permet en plus de traiter couramment des corpus de très grande taille.

Références

Abney, Stephen P (1994). *Parsing by Chunks.* Bell Communication Research

Brill, Eric (1992). *A Simple Rule Based Part of Speech Tagger.* ACL

Dunning, Ted.Dan (1993). *Accurate Methods for the Statistics*. Computational Linguistics, 19(1): 61- 74

El Hadi, Widad Mustafa. Timimi, Ismaïl. Béguin, Annette. De Brito, Marcilio (2001). *The ARC A3 Project: Terminology Acquisition Tools: Evaluation Method ans Task.* ELDS '01

Frantzi, Katerina T. Ananiadou, Sophia. Tsujii, Junichi (1998). *The C-value/NC-value Method of Automatic Recognition of Multi-word Terms.* ECDL'98 pp 585-604

Heinecke, Johannes. Smits, Grégory. Chardenon, Christine. Guimier De Neef, Emilie. Maillebuau, Estelle. Boualem, Malek (2008). *TiLT : plate-forme pour le traitement automatique des langues naturelles.* TAL Vol.49

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

Namer, Fiammetta (2000). *Flemm: Un analyseur Flexionnel de Français à base de règles*. Traitement automatique des Langues pour la recherche d'information, Christian Jacquemin (éds). Paris: Hermes, pp.523-47

Nazarenko, Adeline. Zargayouna, Haïfa. Hamon, Olivier. Van Puymbrouck, Jonathan (2009). *Evaluation des outils terminologiques : enjeux, difficultés et propositions.*TA Vol. 50 pp 257-281

Timimi, Ismaïl (2006). *Evaluation des systèmes d'acquisition de terminologie : nouvelles pratiques, nouvelles métriques.* JADT 2006

Summary

The article describes the evaluation of two terminology extraction tools, Quezao and Acabit. Acabit is well known and freely available, whereas Quezao is an Orange Labs internal product. After a comparison of the theoretical approaches of the two systems, a corpus of news (2424actu) and a corpus of press are used to benchmark the 2 systems (for the qualitative and quantitative aspects).