# Extraction d'entités dans des collections évolutives

Thierry Despeyroux, Eduardo Fraschini, Anne-Marie Vercoustre

▶ **To cite this version:**

Thierry Despeyroux, Eduardo Fraschini, Anne-Marie Vercoustre. Extraction d'entités dans des collections évolutives. 7ièmes Journées francophones Extraction et Gestion des Connaissances EGC 2007, Jan 2007, Namur, Belgique. inria-00116910v2

**HAL Id: inria-00116910**

**https://inria.hal.science/inria-00116910v2**

Submitted on 19 Jun 2007 (v2), last revised 20 Jul 2007 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

inria-00116910, version 2 - 19 Jun 2007

# Extraction d'entités dans des collections évolutives

Thierry Despeyroux*, Eduardo Fraschini*
Anne-Marie Vercoustre*

INRIA- Rocquencourt
Domaine de Voluceau
B.P.105 - 78153 Le chesnay Cedex
* prenom.nom@inria.fr
http://www-rocq.inria.fr/who/Prenom.Nom/

**Résumé.** Le but de notre travail est d'exploiter un ensemble de rapports pour en extraire des entités nommées, en l'occurence des listes de partenaires. À partir d'une liste initiale, nous utilisons un premier ensemble de documents pour identifier des schémas syntaxiques qui sont ensuite validés par apprentissage supervisé sur des documents annotés pour en mesurer l'efficacité avant d'être utilisés sur l'ensemble des documents à explorer. Cette approche est inspirée de celle utilisée pour l'extraction de données dans les documents semi-structurés (wrappers) et ne nécessite pas de ressources linguistiques particulières ni de larges collections de tests. Notre collection de documents évoluant annuellement, nous espérons une amélioration de notre extraction dans le temps.

## 1 Introduction

Le schéma le plus classique en recherche d'information est le suivant : à partir d'une question, d'un ensemble de mots clés, d'un sujet, trouver les documents qui répondent à cette question, qui contiennent ces mots clés, qui parlent du sujet.

Dans notre travail, la recherche est inverse puisque nous voulons extraire dans une collection donnée de documents un certain type d'information, en l'occurrence des entités nommées. Nous devons donc chercher dans des textes une information d'un type donné.

Ce travail d'extraction se fait généralement dans le but d'annoter des documents avec des informations sémantiques qui permettent ensuite une indexation et un accès à l'information de plus haut niveau qu'en utilisant le texte brut. Ces entités sont souvent des noms de personnes, d'organisation, des dates, évènements, lieux etc. Pour mener à bien l'extraction, on utilise généralement des dictionnaires, thésaurus, des listes d'entitées connues, des algorithmes de pattern-matching, des méthodes d'apprentissage en général couteuses sur de larges corpus de documents.

La Wikipedia nous apprend que la reconnaissance d'entités nommées (appelée aussi identification d'entités) et une sous discipline de l'extraction d'information qui cherche à localiser et classer des éléments atomiques d'un texte en catégories prédéfinies telles que des noms de personnes, organisations, localisation, dates, quantités, valeurs monétaires, pourcentages etc. Ce domaine de recherche est très actif, bien que des outils commerciaux existent déjà.

Parmi ces outils, sitons REX (Rosette® Entity Extractor) (http ://www.basistech.com/entity-extraction/), Inxight SmartDiscovery (http ://www.inxight.com/products/smartdiscovery/ee/), Convera - RetrievalWare Entity Extraction and Xeros-Research Entity Extraction system. Tous ces systèmes ont pour but de localiser des termes génériques tels que "Vice President" et "earnings estimates", des dates et éventuellement de les fournir pour des traitements ultérieurs sous forme de méta-data. Les techniques diffèrent et peuvent faire appel à des listes d'entités connues, des algorithmes de recherche de shémas, des méthodes statistiques qui demandent de l'apprentissage sur de grosses quatités de données déjà annotées, des règles d'extraction lexicales, contextuelles ou syntaxiques données manuellement ou inférées par apprentissage.

Au point de vue recherche, des systèmes de reconnaissance d'entités nommées utilisant des techniques à base de grammaires linguistiques ou des modèles statistiques ont été créés. Les systèmes à base de grammaires contruits à la main obtiennent de bien meilleurs résultats au prix d'un travail par des linguistes chévronnées très important. En revanche, les systèmes à base de modèles statistiques demandent beaucoup de données d'apprentissage annotées, mais sont plus faciles à porter vers d'autres langages, domaines ou genre de textes.

Dans notre application pratique, il s'agit plus concrètement de retrouver dans un ensemble de rapports d'activité annuels (ceux de l'Inria), une liste de partenaires avec qui les équipent de recherche coopèrent (contrats, relations internationales, etc.). Dans ce contexte, les mesures de "recall" deviennent donc plus importantes que dans d'autres contextes puisque que le résultat peut être revu manuellement, même si cette tâche de vérification s'avère très lourde pratiquement. D'autre part, ce genre de rapport étant répétitif d'une année sur l'autre, il est intéressant que le processus d'extraction puisse s'affiner avec le temps.

Nous proposons une approche dans laquelle, à partir d'une liste connue d'entité, le système génère automatiquement des schémas de phrases pouvant contenir ces entités. Une étape d'apprentissages, à partir d'un très petit nombre de document permet de ne garder que les schémas les plus pertinent. Cette approche s'inspire de celle utilisée pour l'extraction de données dans des documents semi-structurés tels que des pages web (wrappers), basée sur la génération de programmes d'extraction à partir d'un petit nombre d'exemples. Au lieu de s'appuyer sur les balises html des documents, nos règles s'appuient sur les syntagmes du langage (balises linguistiques). Cette approche ne nécessite pas de ressources linguistiques particulières (McNamee et Mayfield (2002)) ni de larges collections de tests.

Dans la section suivante nous nous attardons un peu sur les données que nous avons à étudier, avant de présenter en détail notre méthode et nos premiers résultats.

## 2 Domaine applicatif

La matière sur laquelle nous avons travaillé est le rapport d'activité scientifique annuel de l'Inria. Composé d'environ 180 rapports provenant des diverses équipes de recherche ce document est une mine de renseignements sur l'activité scientique de l'institut, mais se pose la question de savoir comment exploiter cette source à des fins à la fois de vérification et de synthèse. Nous nous intéressons ici aux parties parlant des collaborations et des contrats dans le but de construire une liste de partenaires.

Ce travail se heurte à plusieurs difficultés inhérentes à la collection. Le style est très peu homogène, parfois télégraphique ou peu rédigé, avec une représentation des noms de partenaires parfois approximative, voir avec des orthographes éronnées. Ces noms eux-même peuvent être

très divers, avec des sigles plus ou moins développés, des localisations intégrées au nom. Ce peut être des noms d'organismes, de laboratoires, de réseaux et peuvent parfois être confondus avec des noms communs. Enfin, à la fois le travail d'annotation des documents pour la phase d'apprentissage qui permet le réglage des paramètre et la vérification des entités extraites sur la totalité de la collection sont des activités très lourdes et très coûteuses en temps.

Dans un premier temps, nous avons essayé d'utiliser un outil existant, à savoir xxx pour essayer d'extraire des noms de partenaires. Ce premier essai a été assez décevant. Une des raisons est sans doute qu'une analyse syntaxique indépendante du contexte n'est pas suffisante pour identifier des noms d'organismes. En plus de travailler sur les syntagmes, nous allons donc essayer de tenir compte du contexte en identifiant des schémas qui contiennent des noms d'organismes.

# 3 Méthode utilisée

Nous présentons dans cette section l'algorithme que nous avons utilisé pour extraire des rapports d'activité de l'Inria la liste des organismes partenaires cités.

Pour celà nous travaillons sur une version des documents qui ont été au préalable annotés à l'aide d'un analyseur qui détecte la fonction grammaticale (nom, verbe etc.) des mots utilisés. Ce sont ces fonction grammaticales (syntagmes) qui seront utilisés pour la construction de schémas tel que nous en parlerons plus loin.

L'ensemble des rapports d'activités n'est pas utilisé dans ça totalité dès le départ. Nous travaillons tout d'abord sur un ensemble réduit qui permet à la fois à notre algorithme "d'apprendre" et aussi d'évaluer sa performance car nous l'auront complètement analysé à la main auparavant.

Cet ensemble réduit est divisé en trois sous ensembles. Le premier, ensemble initial $L$, servira à construire une liste initiale de noms d'organisme ainsi qu'une liste initiale de schémas textuels contenant ces noms. Le deuxième sous-ensemble, $A$, servira comme base d'apprentissage. Le troisième, $B$, comme base de test.

## 3.1 Les schémas

Tous nos textes sont annotés (taggés) par l'analyseur ANNIE. Les balises indiquent les catégories syntaxiques (par ex. NNP pour nom propre, NNPS pour nom pluriel, IN pour préposition) et précèdent les mots annotés. Nous avons ajouté la paire de balises `<org>, </org>` indique les noms d'organisme. Elles sont ajoutés manuellement lors de la préparation des ensembles $L$, $A$ et $B$ et automatiquement pendant l'exécution de l'algorithme d'apprentissage.

Ainsi, le fragment de phrase "by Texas Instrument because" devient `<cat=″in>by <org><cat=″nnp″>Texas <cat=″nnps″>Instruments </org><cat=″in>because`. L'algorithme expliqué plus loin va générer à partir de cette phrase le schémas `IN NNPS*une nouvelle catégorie qui fusionne NNP et NNPS. Le signe  indique une répétition arbitraire du syntagmes qui précèdent. L'algorithme qui suis génère des schémas poru tous les organismes trouvés, avec un contexte gauche et droit de longueur compris entre 1 et 5. Ces schémas sont très généraux puisque seule la catégorie syntaxique est retenue, et non la valeur (par ex. NNP et non pas University).`

## 3.2 L'algorithme

Nous présentons maintenant notre méthode avec plus de détails.

1- L'ensemble des documents L, A et B est analysé et marqué à la main.

2- À partir de la liste d'organismes contenu par l'ensemble de documents initial, nous construisons en explorant l'ensemble initial ainsi que l'ensemble d'apprentissage $(L+A)$ une liste de schémas $S_{LA}$ qui contiennent ces noms d'organisme. La question est de savoir si schémas sont pertinents pour identifier des noms d'organismes. En appliquant ces schémas à l'ensemble initial ainsi qu'à l'ensemble d'apprentissage nous extrayons ensuite une liste d'organismes potentiels. Pour chaque schémas nous pouvons calculer combien de fois il a détecté correctement une organisation ($C$)et combien de fois il a échoué ($E$) et éliminer certains schémas à partir de seuils ($C/Card(L+A) < C_0$ et $E/Card(L+A) > E_0$), puis classer ceux qui restent par niveau de performance ($C/E$).

3- En partant de la liste connue des organismes de la liste initiale il s'agit d'ajouter un à un les schémas classés précédemment, en partant du plus performant et à extraire à chaque fois la liste des organismes dans l'ensemble initial et l'ensemble d'apprentissage. En principe, rappel et précision vont augmenter puisque les schémas les plus performants sont ajoutés en premier. Quand le rappel et la précision se dégradent et passe en dessous d'un seuil fixé, nous arrêtons l'itération car celà signifie que nous utilisont des règles pas assez performantes.

```
PSet= {}; ResList = org(L);
do {A = TagWithList(ResList,A);
    PSet += top(S_LA); S_LA = pop(S_LA);
    A = TagWithPatterns(PSet, A);
    ResList += ExtractOrgs(A);  }
while Recall(ResList, Orgs(A)) < MinRecall &
      Precision(ResList, Orgs(A)) > MaxPrecision
```

4- A cette étape du processus, nous avons donc pu sélectionner des schémas d'extraction dont nous avons pu contrôler la performance sur l'ensemble initial et sur l'ensemble d'apprentissage. Il reste maintenant à utiliser les règles sélectionnéees par ce processus pour rechercher dnas l'ensemble de test les noms d'organisation. Comme l'ensemble de test a été lui aussi analysé au préalable, nous pouvons de nouveau controler la précision et le rappel pour valider notre approche.

| | L | A | B | total |
|---|---|---|---|---|
| Documents | 4 | 8 | 8 | 20 |
| Entités différentes | ? | ? | ? | ? |
| Occurences d'entités | 118 | 218 | 352 | 688 |

**TAB. 1** – *Nombre de documents et d'entités dans la collection initiale utilisée pour l'apprentissage et les tests*

# 4 Expériences et résultats

Nous avons utilisé 20 documents choisis au hasard pour lesquels nous avons identifié à la main la liste des organismes cités (entités). Cet ensemble de document a été divisé en trois ensembles (L, A, B), où L est utilisée pour déterminer la liste initiale[1] d'organismes, l'ensemble A pour l'apprentissage des schémas pertinents, et l'ensemble B est utilisé comme ensemble test pour la validation. Les valeurs des données sont résumées dans le tableau 1.

Le marquage manuel des documents est un travail long et difficile, car les organismes ne sont généralement pas connus. Nous avons décidés de conserver les noms de projets de recherche de l'Inria dans les noms d'entités acceptables, car ce sont souvent des acronymes et ils apparaissent dans les phrases exactement comme des noms d'organismes. Puisque la liste officielle des projets est connue, il ne serait pas difficile de les éliminer si on ne souhaite pas les conserver (encore qu'on peut être intéressé à savoir quelle équipe travaille avec telle autre).

Comme il est extrêmement fastidieux et difficile d'identifier les organismes cités dans les documents, nous ne voulions pas avoir à le faire pour plus de 20 documents. Afin de tester différents paramètres de l'algorithme, nous avons effectué des permutations aléatoires des documents dans les ensembles L, A et B afin de créer 10 jeux de tests différents.

Le tableau 2 présente les résultats pour 3 de ces jeux de test, pour des seuils d'apprentissage de 0,7 pour la précision et 0, 4 pour le rappel. Nous nous intéressons à la fois à l'identification des noms d'organismes (comptage simple), et à l'identification des occurrences de ces noms (comptage multiple).

On peut tout d'abord remarquer que le rappel de départ pour l'ensemble d'apprentissage A est faible, ce qui indique une grande diversité de partenaires selon les différentes équipes. Le rappel à la fin de la période d'apprentissage a été multiplié par plus de 2 en moyenne pour le comptage simple, mais reste malgré tout assez faible. Il

[1] Cette liste aurait pu être déterminée par une méthode quelconque, par exemple la liste des partenaires officiels de l'Inria trouvée sur le site de l'Inria (principalement français), augmentée d'organismes choisis a priori.

| | Expérience 1 | Expérience 2 | Expérience 3 |
|---|---|---|---|
| Entités dans L | 77 | 70 | 63 |
| Entités de L trouvées dans A | ? | ? | ? |
| Occurrences d’entités de L trouvées dans A | 49<br>(R=0,15 ; P=0,76)<br>(MR=0,19 ; MP=0,76) | 120<br>(R=0,11 ; P=0,48)<br>(MR=0,15 ; MP=0,51) | 104<br>(R=0,23 ; P=0,92)<br>(MR=0,31 ; MP=0,89) |
| Schémas retenus | 66 | 81 | 17 |
| Entités dans A à la fin de l’apprentisage | 121<br>(R=0,35 ; P=0,69)<br>(MR=0,39 ; MP=0,70) | 178<br>(R=0,33 ; P=0,51)<br>(MR=0,40 ; MP=0,51) | 96<br>(R=0,36 ; P=0,87)<br>(MR=0,42 ; MP=0,84) |
| Entités de A trouvées dans B | 112/121 | 154/178 | 55/96 |
| Entités extraites de B | 222 | 228 | 107 |
| Comptage simple | R=0,36 ; P=0,52 | R=0,42 ;P=0,40 | R=0,19 ; P=0,87 |
| Comptage multiple | R=0,42 ; P=0,37 | R=0,47 ;P=0,37 | R=0,22 ; P=0,71 |

TAB. 2 – *titre...*

faut rappeler que l’algorithme d’apprentissage s’arrête lorsque le rappel est plus grand que 0.4 ou la précision inférieure à 0.6 (pour le comptage multiple???). En particulier, dans l’expérience 3, le recall passe assez vite au-dessus de ce seuil, ce qui explique le petit nombre de schémas sélectionnés. On ne peut donc pas espérer des valeurs très élevées à la fin de l’apprentissage, en utilisant des schémas génériques et calculés automatiquement.

Le rappel final sur l’ensemble test B est meilleur que sur l’ensemble d’apprentissage, ce qui pourrait indiquer que l’algorithme d’extraction marche d’autant mieux que la liste de départ utilisée est grande. (à vérifier).

2) Nous avons voulu tester l’influence de ces seuils sur les résultats.

Resultats?

3) les organismes partenaires mentionnés sont de nature très différente. En effet, certains partenaires sont des partenaires avec lesquels nous avons des contrats explicites, ou qui sont impliqués avec nous dans des projets nationaux ou internationaux. Par ailleurs nous avons beaucoup de collaborations plus académiques avec des universités et instituts français ou étrangers. Les premiers sont plutôt mentionnés dans la section "contrats" des rapports, et les second dans la section "collaborations". Nous avons donc refaits les expériences en séparant les sections "contrats" et les sections "collaborations" pour évaluer 1) si un type d’organisme était plus difficile à identifier que l’autre 2) si un apprentissage plus spécifique sur chacun des groupes améliorait les résultats.

## 5 Conclusion

D'une année sur l'autre, il y a une certaine continuité dans les partenaires avec lesquels les équipes INRIA travaillent. Il est donc raisonnable d'utiliser la liste des organisations d'une année pour initialiser l'extraction d'entités pour l'année N+1. Cette liste peut facilement être nettoyée à la main, de même que la liste finale. C'est certainement beaucoup plus rapide que d'extraire manuellement le nom des organismes à partir des 150 rapports d'activité. On est donc plus intéressé par un bon rappel qu'une bonne précision. Cela a joué dans le choix des seuils utilisés dans l'apprentissage.

## Références

McNamee, P. et J. Mayfield (2002). Entity extraction without language-specific resources.

## Summary

The goal of our work is to use a set of reports and extract named entities, in our case the names of partners. Starting with an initial list of entities, we use a first set of documents to identify syntactic patterns that are then validated in a supervised learning phase on a set of annotated documents to perform a performance test. The complete collection is then explored. This approach comes from the one that is used in data extraction for semi-structured documents (wrappers) and do not need any linguistic ressources neither a large set for training. As our collection of documents evoluate, we hope that the performance of the extraction becomes better year after year.