Découverte de connaissances dans les séquences par CRF non-supervisés

Vincent Claveau[1] Abir Ncibi[2]
(1) IRISA-CNRS (2) INRIA-IRISA
Campus de Beaulieu, 35042 Rennes, France
vincent.claveau@irisa.fr abir.ncibi@inria.fr

RÉSUMÉ

Les tâches de découverte de connaissances ont pour but de faire émerger des groupes d'entités cohérents. Ils reposent le plus souvent sur du clustering, tout l'enjeu étant de définir une notion de similarité pertinentes entre ces entités. Dans cet article, nous proposons de détourner les champs aléatoires conditionnels (CRF), qui ont montré leur intérêt pour des tâches d'étiquetage supervisées, pour calculer indirectement ces similarités sur des séquences de textes. Pour cela, nous générons des problèmes d'étiquetage factices sur les données à traiter pour faire apparaître des régularités dans les étiquetages des entités. Nous décrivons comment ce cadre peut être mis en œuvre et l'expérimentons sur deux tâches d'extraction d'informations. Les résultats obtenus démontrent l'intérêt de cette approche non-supervisée, qui ouvre de nombreuses pistes pour le calcul de similarités dans des espaces de représentations complexes de séquences.

ABSTRACT

Unsupervised CRF for knowledge discovery

Knowledge discovery aims at bringing out coherent groups of entities. They are usually based on clustering ; the challenge is then to define a notion of similarity between the relevant entities. In this paper, we propose to divert Conditional Random Fields (CRF), which have shown their interest in supervised labeling tasks, in order tocalculate indirectly the similarities among text sequences. Our approach consists in generate artificial labeling problems on the data to be processed to reveal regularities in the labeling of the entities. We describe how this framework can be implemented and experiment it on two information retrieval tasks. The results demonstrate the usefulness of this unsupervised approach, which opens many avenues for defining similarities for complex representations of sequential data.

MOTS-CLÉS : Découverte de connaissances, CRF, clustering, apprentissage non-supervisé, extraction d'informations.

KEYWORDS: Knowledge discovery, CRF, clustering, unsupervised machine learning, information extraction.

1 Introduction

Les tâches d'étiquetage de séquences sont depuis longtemps d'un intérêt particulier pour le TAL (étiquetage en parties-du-discours, annotation sémantique, extraction d'information, etc.). Beaucoup d'outils ont été proposés pour ce faire, mais depuis quelques années, les Champs aléatoires conditionnels (*Conditional Random Fields*, CRF (Lafferty *et al.*, 2001)) se sont imposés comme l'un des plus efficaces pour de nombreuses tâches. Ces modèles sont supervisés : des exemples de séquences avec leurs labels sont donc nécessaires.

Le travail présenté dans cet article se place dans un cadre différent dans lequel on souhaite faire émerger des informations à partir de ces séquences. Nous nous inscrivons donc dans une tâche de découverte de connaissances dans laquelle il n'est plus question de supervision, le but étant au contraire de découvrir comment les données peuvent être regroupées dans des catégories qui fassent sens. Ces tâches de découvertes reposent donc le plus souvent sur du clustering (Wang *et al.*, 2011, 2012; Ebadat *et al.*, 2012), la question cruciale étant de savoir comment calculer la similarité entre deux entités jugées intéressantes. Dans cet article, nous proposons de détourner les CRF en produisant des problèmes d'étiquetage factices pour faire apparaître des entités régulièrement étiquetées de la même façon. De ces régularités est alors tirée une notion de similarité entre les entités, qui est donc définie par extension et non par intention.

D'un point de vue applicatif, outre l'usage pour la découverte de connaissances, les similarités obtenues par CRF et le clustering qu'il permet peut servir en amont de tâches supervisées :
– il peut être utilisé pour réduire le coût de l'annotation de données. Il est en effet plus simple d'étiqueter un cluster que d'annoter un texte instance par instance.
– il peut permettre de repérer des classes difficiles à discerner, ou au contraire d'exhiber des classes dont les instances sont très diverses. Cela permet alors d'adapter la tâche de classification supervisée en modifiant le jeu d'étiquettes.

Dans la suite de cet article, nous positionnons notre travail par rapport aux travaux existants et présentons brièvement les CRF en introduisant quelques notions utiles pour la suite de l'article. Notre décrivons ensuite en section 3 le principe de notre approche de découverte utilisant les CRF en mode non-supervisé pour faire de la découverte dans des séquences. Nous proposons deux expérimentations de cette approche dans les sections 4 et 5, puis nous présentons nos conclusions et quelques pistes ouvertes par ce travail.

2 Travaux connexes

Comme nous l'avons mentionné en introduction, les tâches d'étiquetage de séquences sont très courantes en traitement automatique des langues. Celles-ci se présentent souvent dans un cadre supervisé, c'est-à-dire que l'on dispose de séquences annotées par des experts, et incidemment du jeu de label à utiliser. C'est dans ce cadre que les CRF se sont imposés comme des techniques d'apprentissage très performantes, obtenant d'excellents résultats pour de nombreuses tâches (Wang *et al.*, 2006; Pranjal *et al.*, 2006; Constant *et al.*, 2011; Raymond et Fayolle, 2010, entre autres).

Plusieurs études ont proposé de passer à un cadre non-supervisé. Certaines ne relèvent pas à proprement parler de non-supervision mais plutôt de semi-supervision, où le but est de limiter le nombre de séquences à annoter. C'est notamment le cas pour la reconnaissance d'entités nommées

où beaucoup de travaux s'appuient sur des bases de connaissances extérieures (Wikipedia par exemple), ou sur des règles d'extraction d'amorçage données par un expert (Kozareva, 2006; Kazama et Torisawa, 2007; Wenhui Liao, 2009; Elsner *et al.*, 2009). On peut également citer les travaux sur l'étiquetage en parties du discours sans données annotées (Merialdo, 1994; Ravi et Knight, 2009; Richard et Benoit, 2010). Dans tous les cas, l'angle de vue de ces travaux est la limitation, voire la suppression, des données d'apprentissage. Ils ne se posent pas dans un cadre de découverte de connaissances : ils reposent donc sur un *tagset* déjà établi, même si la correspondance mot-tag peut n'être qu'incomplètement disponible (Smith et Eisner, 2005; Goldwater et Griffiths, 2007).

Le cadre que nous adoptons dans cet article est différent puisque nous proposons de faire émerger les catégories de données non annotées. À l'inverse des travaux précédents, nous ne faisons donc pas d'a priori sur les étiquettes possibles. Notre tâche relève donc d'un clustering dans lequel les éléments similaires des séquences doivent être groupés, comme cela a été fait par exemple par Ebadat *et al.* (2012) pour certaines entités nommées. Le clustering de mots n'est pas une tâche nouvelle en soi, mais elle repose sur la définition d'une représentation pour les mots (typiquement un vecteur de contexte) et une mesure de distance (ou de similarité, typiquement un cosinus). Notre approche a pour but d'utiliser la puissance discriminative des CRF, qui a montré son intérêt dans le cas supervisé, pour offrir une mesure de similarité plus performante. Il s'agit donc de transformer cette technique supervisée en méthode non-supervisée permettant de déterminer la similarité entre deux objets.

Ce détournement de techniques d'apprentissage supervisé pour faire émerger des similarités dans des données complexes non étiquetées a déjà été utilisé. Il a montré son intérêt sur des données de type attributs-valeurs pour lesquelles la définition d'une similarité était difficile (attributs non numériques, biais d'une définition ex nihilo), notamment avec le *random forest clustering* (Liu *et al.*, 2000; Hastie *et al.*, 2001). L'approche consiste à générer un grand nombre de problèmes d'apprentissage factices, avec des données synthétiques mélangées aux données réelles, et de voir quelles données sont classées régulièrement ensemble (Shi et Horvath, 2005). Notre approche s'inscrit dans ce cadre, mais exploite les particularités des CRF pour pouvoir prendre en compte la nature séquentielle de nos données.

2.1 Champs aléatoires conditionnels

Les CRF (Lafferty *et al.*, 2001) sont des modèles graphiques non dirigés qui cherchent à représenter la distribution de probabilités d'annotations (ou étiquettes ou labels) y conditionnellement aux observations x à partir d'exemples labellisés (exemples avec les labels attendus). Ce sont donc des modèles obtenus par apprentissage supervisé, très utilisés notamment dans les problèmes d'étiquetage de séquences. Un bonne présentation des CRF peut être trouvée dans ? ? ?. Nous ne présentons ci-dessous que les éléments et notations utiles pour la suite de cet article.

Dans le cas séquentiel, c'est-à-dire l'étiquetage d'observations x_i par des labels y_i, la fonction potentielle au cœur des CRF s'écrit :

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_{k=1}^{k_1}\sum_{i=1}^{n}\lambda_k f_k(y_i, x) + \sum_{k=1}^{k_2}\sum_{i=1}^{n}\mu_k g_k(y_{i-1}, y_i, x)\right) \quad (1)$$

avec :

– $Z(x)$ un facteur de normalisation ;
– les fonctions caractéristiques locales et globales (fonctions *features*) f et g : les fonctions f caractérisent les relations locales entre le label courant en position i et les observations ; les fonctions g caractérisent les transitions entre les nœuds du graphe, c'est-à-dire entre chaque paires de labels i et $i-1$, et la séquence d'observations.
– les valeurs k_1, k_2 et n sont respectivement le nombre de fonctions *features* f, le nombre de fonctions *features* g, et la taille de la séquence de labels à prédire.

Les fonctions f et g sont généralement des fonctions binaires vérifiant une certaine combinaison de labels et d'attributs décrivant les observations et appliquées à chaque position de la séquence. Ces fonctions sont définies par l'utilisateur ; elles reflètent sa connaissance de l'application. Elles sont pondérées par les λ_k et μ_k qui estiment l'importance de l'information qu'elles apportent pour déterminer la classe.

L'apprentissage des CRF consiste à estimer le vecteur de paramètres $\theta = \lambda_1, \lambda_2, ..., \lambda_{k_1}, \mu_1, \mu_2, ..., \mu_{k_2}$ (poids des fonctions f et g) à partir de données d'entraînement, c'est-à-dire N séquences étiquetées $(x^{(i)}, y^{(i)})_{i=1}^{i=N}$. en pratique, ce problème est ramené à un problème d'optimisation, généralement résolu en utilisant des méthodes de type quasi-Newton, comme l'algorithme L-BFGS (Schraudolph *et al.*, 2007). Après cette étape d'apprentissage, l'application des CRF à de nouvelles données consiste à trouver la séquence de labels la plus probable étant donnée une séquence d'observations non-vue. Comme pour les autres méthodes stochastiques, celle-ci est généralement obtenu avec un algorithme de Viterbi.

3 Principes du modèle non supervisé

Nous décrivons dans cette section le principe de notre approche. Une vue générale est tout d'abord donnée au travers d'un algorithme schématisant l'ensemble du processus. Nous en détaillons ensuite quelques points cruciaux, ainsi que des aspects plus pragmatiques de l'utilisation de cette méthode.

3.1 Principe général

Comme nous l'avons expliqué précédemment, l'idée principale de notre approche est de déduire une distance (ou une similarité) à partir de classifications répétées de deux objets pour des tâches d'apprentissage aléatoire. Plus les objets sont détectés souvent comme appartenant à la même classe, plus ils sont supposés proches. L'algorithme 1 donne un aperçu global de la démarche. Dans notre cadre séquentiel, la classification est faite grâce aux CRF (les étapes 6 et 7 correspondent simplement à l'apprentissage et l'application d'un modèle CRF). Celle-ci est répétée un grand nombre de fois en faisant varier les données, les labels (les ω_i sont des classes factices) et les paramètres des apprentissages. Il est tenu à jour un compte des paire de mots (x_i, x_j) recevant les mêmes labels ; ces co-étiquetages sont contenus dans la matrice $\mathscr{M}_{\text{co-et}}$. Ils sont mis à jour à chaque itération en tenant compte éventuellement de différents critères, selon une fonction weight (cf. infra pour une discussion sur ce point). Ces co-étiquetages sont ensuite transformés en mesures de similarité (cela peut être une simple normalisation) collectées dans $\mathscr{M}_{\text{sim}}$.

Algorithme 1 Clustering par CRF

input : $\mathscr{E}_{\text{total}}$: séquences non étiquetées
for grand nombre d'itérations **do**
  $\mathscr{E}_{\text{train}}$, $\mathscr{E}_{\text{app}} \leftarrow$ Diviser($\mathscr{E}_{\text{total}}$)
  Tirer aléatoirement les labels y_i parmi $\omega_1 ... \omega_L$ pour les séquences de $\mathscr{E}_{\text{train}}$
  Générer aléatoirement un ensemble de fonctions f et g
  Inférence : $\theta \leftarrow$ L-BFGS($\mathscr{E}_{\text{train}}$,$y$,$f$,$g$)
  Application : $y^* = \arg\max_y p_{(\theta,f,g)}(y|x)$ pour tous les $x \in \mathscr{E}_{\text{app}}$
  for all classe ω_l parmi $\omega_1 ... \omega_L$ **do**
    for all paire x_i, x_j de $\mathscr{E}_{\text{app}}$ telle que $y_i^* = y_j^* = \omega_l$ **do**
      $\mathscr{M}_{\text{co-et}}(x_i, x_j) += \text{weight}(x_i, x_j, \omega_l)$
    end for
  end for
end for
$\mathscr{M}_{\text{sim}} \leftarrow$ Transformation($\mathscr{M}_{\text{co-et}}$)
$\mathscr{C}_{CRF} \leftarrow$ Clustering($\mathscr{M}_{\text{sim}}$)
return $\mathscr{C}_{CRF}$

3.2 Apprentissage aléatoire

L'approche repose sur le fait que les CRF vont permettre d'exhiber une similarité entre des mots en leur attribuant régulièrement les mêmes étiquettes dans des conditions d'apprentissage très variées. Pour cela, à chaque itération, plusieurs choix aléatoires sont mis-en-œuvre ; ils concernent :
– les séquences servant à l'apprentissage et leur nombre ;
– les labels (distribution et nombre) ;
– les fonctions *features* décrivant les mots ;
Ces apprentissages sur des tâches supervisées factices doivent ainsi conférer, par leur variété, des propriétés importantes à la similarité obtenue. Celle-ci mélange ainsi naturellement des descriptions complexes (attributs nominaux divers sur le mot courant, sur les mots voisins), opère par construction une sélection de variables et prend ainsi en compte les redondances des descripteurs ou ignore ceux de mauvaise qualité, et elle est robuste aux données aberrantes.

Bien sûr, comme nous l'avons déjà souligné, ce rôle important de l'aléatoire n'empêche pas l'utilisateur de contrôler la tâche via des biais. Cela se traduit par exemple par la mise à disposition des descriptions riches des mots : étiqueter des séquences en parties-du-discours, apport d'informations sémantiques sur certains mots... Cela se traduit également par la définition de l'ensemble des fonctions *features* parmi lesquelles l'algorithme peut piocher les fonctions f et g à chaque itération. Dans les expériences rapportées ci-dessous, cet ensemble de fonctions est celui classiquement utilisés en reconnaissance d'entités nommées : forme et parties du discours du mot courant, des 3 précédents et 3 suivants, des bigrammes de ces attributs, casse des mots-formes courants et environnants... Concernant les ensembles $\mathscr{E}_{\text{train}}$ et $\mathscr{E}_{\text{app}}$, à chaque itération 5 % des phrases sont tirées aléatoirement pour constituer l'ensemble d'entraînement ; le reste sert d'ensemble d'application.

3.3 Labels aléatoires

Le choix du nombre de labels factices et leur distribution est également important (mais il faut noter que le nombre de labels choisi à ce stade n'implique pas directement le nombre de clusters qui seront produits lors de l'étape finale de clustering). Un trop grand nombre de labels lors de l'apprentissage risque d'empêcher de produire ensuite un étiquetage dans lequel peu d'entités partagent le même label. En soit ce problème ne pose pas nécessairement un problème de qualité finale, mais risque d'augmenter le nombre d'itérations suffisant pour l'obtention de ce résultat final. À l'inverse, si l'on choisit un nombre trop restreint de labels, l'application du modèle risque de de ne pas suffisamment différencier les entités, produisant des co-étiquetages fortuits. Ce problème est plus gênant car il va impacter le résultat du *clustering*. Il faut de plus noter que tout cela est à interpréter selon les autres paramètres de l'apprentissage. Ainsi, les fonctions *features* vont permettre ou pas un sur-apprentissage, et donc éventuellement empêcher ou favoriser les co-étiquetages. La taille de $\mathscr{E}_{\text{train}}$, et notamment le nombre d'entités y recevant un même label intervient aussi : si systématiquement dès l'entraînement un grand nombre d'entités, probablement de classes différentes, reçoivent le même label, les modèles ne vont pas être correctement discriminants.

Pour correctement prendre en compte ce phénomène, il serait nécessaire de caractériser la propension du modèle appris, avant l'étiquetage, à trop ou pas assez discriminer les entités. Dans l'état actuel de nos travaux, nous n'avons pas formalisé un tel critère. Nous utilisons simplement un critère a posteriori déterminé sur le texte après étiquetage : un co-étiquetage de deux entités "rapporte plus" si peu d'entités ont été étiquetées avec ce même label. Cela est mis en œuvre dans la fonction *weight* utilisée pour mettre à jour la matrice $\mathscr{M}_{\text{co-et}}$. En pratique, dans les expériences rapportées dans cet article, on a défini cette fonction par : $\text{weight}(x_i, x_j, \omega_l) = \frac{1}{|\{x_k | y_k = \omega_l\}|}$ et le nombre de labels est lui aussi tiré aléatoirement entre 10 et 50 à chaque itération.

Il est aussi possible, selon le problème traité et les connaissances particulières qui s'y appliquent, de biaiser la distribution des étiquettes aléatoires. Ainsi, pour un problème donné, si l'on sait que toutes les occurrences d'un mot-forme ont forcément la même classe, il est important que cette contrainte soit mise en œuvre lors de la production des données d'entraînement. L'expérience rapportée en section 4 se place dans ce cadre.

3.4 Clustering

L'étape finale de clustering peut être mise en œuvre de différentes façons grâce aux techniques et outils existants. L'algorithme célèbre du *k-means* qui nécessite des calculs de barycentres durant le processus n'est bien sûr pas adapté à notre espace non métrique. Sa variante *k-medoids*, qui utilise un objet comme représentant d'un cluster et ne nécessite donc pas d'autres mesures que celles fournie par $\mathscr{M}_{\text{sim}}$, peut l'être.

Il faut cependant noter que dans nos tâches de découverte, le nombre de clusters attendus est inconnu. Pour notre part, dans les expérimentations présentées dans les sections 4 et 5, nous utilisons donc une autre technique de clustering, le Markov Clustering (MCL). Cette technique a été développée initialement pour le partitionnement de grands graphes (van Dongen, 2000). Son avantage par rapport au *k-medoids* est de ne pas nécessiter de fixer a priori le nombre de clusters attendus, et aussi d'éviter le problème de l'initialisation de ces clusters. Nous considérons donc

simplement nos objets (mots ou autres entités) comme des nœuds d'un graphe dont les arcs sont valués en fonction de la similarité contenue dans $\mathscr{M}_{\text{sim}}$.

3.5 Aspects opérationnels

Appliqué tel quel, le processus exposé en section 3 va considérer tous les éléments composant les séquences et tenter de les organiser en clusters. Dans beaucoup d'applications, la tâche de clustering n'est intéressante que pour une sous-partie de ces éléments. C'est par exemple le cas en reconnaissance d'entités nommées ou plus largement en extraction d'information, où seuls certains mots ou groupe de mots doivent être considérés. Dans ce cadre, il est très courant d'utiliser des labels dits BIO (Begin-In-Out) qui permettent de modéliser le fait qu'une entité soit multi-mot (le B pour Begin identifie le début de l'entité, le I pour In la continuité et le O indique le mot ne fait pas partie de l'entité). Voici un exemple de séquences factices tiré des données utilisées en section 5 :

x	l'	audience	entre	nicolas	sarkozy	et	maître	wade
	DET	NC	PREP	NP	NP	COO	NC	NP
y	O	O	O	B-fake140	I-fake140	O	B-fake25	B-fake3

Cette connaissance externe fait partie des biais indispensables pour cadrer le processus d'apprentissage non-supervisé et faire en sorte qu'il s'applique aux besoins spécifiques de l'utilisateur. Mais il est important de noter que cette connaissance sur les entités à considérer n'est pas de même ordre que celle l'on se propose de découvrir via le clustering. Dans le premier cas, il s'agit de délimiter les entités intéressantes, dans le second cas, il s'agit d'en faire émerger des classes, sans a priori leur nature.

Il est possible dans ce cas de supposer que l'on sait délimiter les entités intéressantes dans les séquences ; c'est l'hypothèse adoptée dans plusieurs travaux sur la classification d'entités nommées (Collins et Singer, 1999; Elsner *et al.*, 2009; Ebadat *et al.*, 2012). Il est aussi, bien sûr, possible de considérer ce problème comme un problème d'apprentissage pour lequel l'utilisateur doit fournir quelques exemples. Dans les deux cas, cela nécessite de l'expertise, fournie soit en intention (critères objectifs pour délimiter les entités), soit en extension (exemples ; cf. sous-section 5.2). Chacune des expériences rapportées ci-dessous adopte l'un de ces cas de figure.

Le processus itératif proposé dans cet article est évidemment coûteux (mais aisément parallélisable). Dans les expériences rapportées ci-après, le nombre d'itérations a été fixé à 1000. Les principales sources de coût en terme de temps de calcul sont l'apprentissage du modèle CRF et son application. Leur complexité est elle-même dépendante de nombreux paramètres, notamment la taille de l'échantillon d'apprentissage, la variété des observations (x), le nombre de classes aléatoires (ω), les attributs considérés (les fonctions *features* f et g)... Pour minimiser l'impact de ce coût, nous utilisons l'implémentation de CRF WAPITI qui optimise les algorithmes standard d'inférence (Lavergne *et al.*, 2010).

4 Validation expérimentale en classification de noms propres

Pour cette première expérience, nous reprenons la problématique et les données de Ebadat *et al.* (2012). Il s'agit de faire émerger les différentes classes de noms propres au sein de résumés de matchs de football. Plus précisément, dans leurs expériences, les auteurs ont cherché à classer

les noms propres à l'échelle du corpus, c'est-à-dire en considérant que toutes les occurrences relevaient de la même entité et donc de la même catégorie. Dans ce jeu de donnée, les entités ne sont donc pas considérées comme possiblement polysémiques ; même si ce point est discutable, il n'est pas remis en cause dans notre expérience pour lequel nous utilisons le jeu de données tel qu'utilisé par Ebadat *et al.* (2012).

4.1 Tâche et données

Le corpus est composé de rapports de matchs minute par minute en français, extraits de différents sites Web. Les événements importants de chaque minute ou presque d'un match y sont décrits (cf. tableau 1) : remplacement de joueurs, fautes, buts...

Minute	Rapport
80	Zigic donne quelques frayeurs à Gallas et consorts en contrôlant un ballon chaud à gauche des 16 mètres au devant du Gunner. Le Valencian se trompe dans son contrôle et la France peut souffler.
82	Changement opéré par Raymond Domenech avec l'entrée d'Alou Diarra à la place de Sidney Govou, pour les dernières minutes. Une manière de colmater les brêches actuelles ?

TABLE 1: Extrait d'un rapport minute-by-minute d'un match de football

Ces données ont été annotées manuellement par des experts selon des classes définies pour répondre à des besoins applicatifs spécifiques (voir Fort et Claveau, 2012). On possède donc une vérité terrain associant à chaque occurrence de chaque nom propre une classe (voir la figure 1a). On remarque sans surprise que ces classes sont très déséquilibrées, avec notamment une classe *joueur* très peuplée.

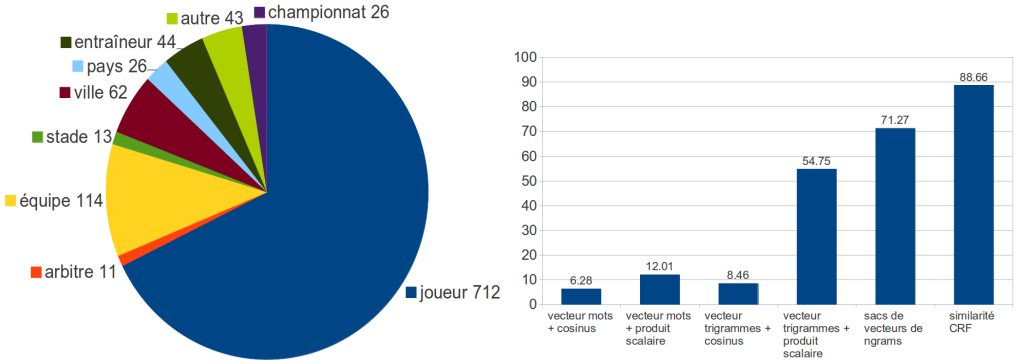
4.2 Mesures de performance

Notre tâche de découverte se ramenant à une étape finale de clustering, nous l'évaluons comme telle. Une telle évaluation est toujours délicate : l'évaluation sur critères externes nécessite de disposer d'un clustering de référence (vérité terrain) dont la pertinence peut toujours être discutée, mais les critères internes (par exemple, une mesure de cohésion des clusters) sont connus pour n'être pas fiables (Manning *et al.*, 2008). Nous nous plaçons donc dans le premier cadre et comparons le clustering obtenu par notre processus à celui de la vérité terrain.

Pour ce faire, différentes métriques ont été proposées, comme la pureté ou *Rand Index* (Rand, 1971). Ces mesures sont cependant peu discriminantes et ont tendance à être trop optimistes quand la vérité terrain contient des classes de tailles très différentes (Nguyen Xuan Vinh, 2010). Nous préférons donc l'*Adjusted Rand Index* (ARI), qui est une version du Rand Index tenant compte des agréments de hasard, et qui est connue pour être robuste. Son étude et sa définition peuvent être trouvées dans (Hubert et Arabie, 1985).

4.3 Implémentation et résultats

Pour tester notre méthode de clustering par CRF, nous avons étiqueté le corpus en partie du discours, utilisé le schéma d'annotation BIO et considéré la phrase comme séquence. Dans cette application particulière, nous reprenons l'hypothèse de (Collins et Singer, 1999; Elsner *et al.*, 2009) : les entités à catégoriser sont connues et délimitées. En pratique, ce sont donc sur elles que les annotations aléatoires vont porter, les autres mots du corpus recevant toujours le même label *'O'*. Les fonctions f et g sont celles classiquement utilisées en extraction d'information : les fonctions f lient le label courant y_i aux observations (forme ou parties du discours du mot courant en x_i, du mot en x_{i-1}, x_{i-2}, x_{i+1}, ou x_{i+2}, ou des combinaisons de ces attributs) ; les fonctions g lient deux labels successifs (y_{i-1}, y_i). La tâche étant de classifier les noms propres au niveau du corpus et non de l'occurrence, nous forçons deux occurrence d'un même nom à avoir le même label lors de la génération des labels aléatoires (étape 4 de l'algorithme). En revanche, l'application du CRF produit une annotation au niveau de l'occurrence, la matrice $\mathcal{M}_{\text{co-et}}$ recense donc les classifications à l'occurrence près. L'étape de transformation (étape 14) permet de transformer cette matrice en une matrice de similarité $\mathcal{M}_{\text{sim}}$ des noms propres au niveau du corpus en sommant les lignes et colonnes des différentes occurrences des mêmes noms.



(a) (b)

FIGURE 1: (a) Répartition des données football dans la vérité terrain (nombre de noms propres uniques). (b) Évaluation des clusterings par rapport à la vérité terrain (ARI %).

Les résultats de notre approche sont donnés dans le tableau 1b en terme d'ARI (en pourcentage ; 0 signifie un clustering aléatoire et 100 un clustering identique à la vérité terrain). À des fins de comparaison, nous reportons les résultats de Ebadat *et al.* (2012) ; ceux-ci ont été obtenus en utilisant des descriptions vectorielles des contextes des entités soit sous forme d'un vecteur unique, soit sous forme de sacs de vecteurs, et des fonctions de similarités adaptées à ces représentations. Le contexte donnant le meilleur résultat est de 4 mots à gauche et à droite de l'entité. L'étape de clustering est faite avec le même algorithme MCL que pour notre système. Ce dernier dispose d'un paramètre d'inflation qui influence indirectement le nombre de cluster produit. Pour une comparaison équitable, les résultats rapportés pour chaque méthode sont ceux pour lesquels ce paramètre est optimal pour la mesure d'évaluation ARI. À titre d'information, cela produit 12 clusters pour la similarité CRF, 11 pour la similarité sac-de-vecteurs n-grams.

Ces résultats soulignent l'intérêt de notre approche par rapport aux représentations et similarités

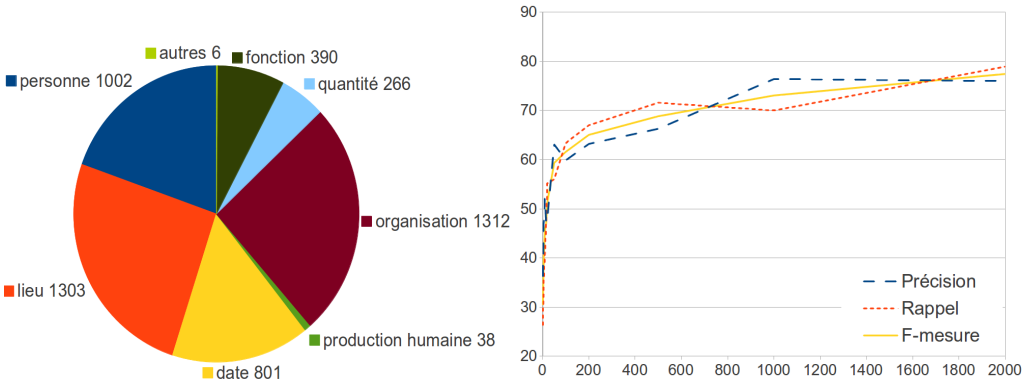
plus standard. Les quelques différences constatées entre les clusters formés par notre approche et les classes de la vérité terrain portent principalement sur la classe *autre*. Celle-ci contient des noms de personnalités apparaissant dans des contextes divers (personnalité donnant le coup d'envoi, apparaissant dans les tribunes...), avec trop peu de régularités pour que les CRF, pas plus que les autres méthodes, arrivent à faire émerger une similarité. Il apparaît également que certaines erreurs rapportées par Ebadat *et al.* (2012) comme récurrentes ne sont pas commises par le clustering par CRF. Par exemple, les méthodes vectorielles ont tendance à confondre les noms de villes et les noms de joueurs, ceux-ci apparaissant souvent proches les uns des autres et partageant donc les mêmes contextes. Ces erreurs ne sont pas commises par l'approche par CRF, où la prise en compte de la séquentialité pour l'étiquetage permet de bien distinguer ces deux classes.

5 Validation expérimentale sur les entités nommées

5.1 Tâche et données

Pour cette tâche, nous utilisons les données de la campagne d'évaluation ESTER2 (Gravier *et al.*, 2005). Elles sont composées de 150h d'émissions de radio datant d'entre 1999 et 2003, provenant de diverses sources (France Inter, Radio Classique, Africa 1...). Ces émissions, transcrites, ont été annotées en entités nommées selon 8 catégories : personnes, fonctions, lieux, organisations, temps, produits humains, quantité, et une catégorie autres.

Contrairement au jeu de données précédent, les entités sont annotées au niveau de l'occurrence et peuvent être des noms propres, communs ou des expressions ; ainsi, l'entité `Paris` peut être annotée comme un lieu ou une organisation selon le contexte. Nous n'utilisons pour nos expériences que la partie *dev* de ce jeu de données ESTER2, transcrite manuellement, mais respectant les particularités d'un système de reconnaissance de la parole : le texte n'a donc ni ponctuation, ni majuscule. Ses caractéristiques sont données dans la figure 2a



(a) (b)

FIGURE 2: (a) Répartition des données ESTER2 dans la vérité terrain (nombre d'occurrences). (b) Performances de la détection des entités selon le nombre de séquences annotées.

5.2 Repérage des entités

Bien qu'il soit possible de se placer dans le même cadre que précédemment et supposer que les entités à classifier sont connues et délimitées, nous utilisons un cadre intermédiaire plus réaliste : nous supposons qu'une petite partie des données est annotée par un expert qui délimite les entités intéressantes (mais sans leur assigner de classe). Ces données vont nous servir dans une première étape à apprendre à délimiter les entités avant de les grouper. On se place donc dans un cadre supervisé classique avec deux classes (entité intéressante ou non), pour lequel nous utilisons les CRF de manière traditionnelle.

La figure 2b présente les résultats obtenus, en fonction du nombre de séquences (phrases) utilisées pour l'apprentissage. Les performances sont évaluées en terme de précision, rappel et F-mesure. Il apparaît qu'il est possible d'obtenir des résultats de bonne qualité en analysant (c'est-à-dire en délimitant les entités nommées) relativement peu de phrases.

5.3 Évaluation du clustering

Nous reprenons le même cadre expérimental que celui expliqué en section 4.3, à la différence que la classification se fait ici au niveau de l'occurrence. La transformation de $\mathscr{M}_{\text{co-et}}$ en $\mathscr{M}_{\text{sim}}$ consiste donc juste en une normalisation. Les entités considérées sont celles repérées par l'étape précédente (avec 2 000 séquences annotées pour l'apprentissage) sur l'ensemble du corpus. Les résultats, mesurés en terme d'ARI (%), sont présentés dans la figure 3. Comme pour l'expérience précédente, nous présentons les résultats obtenus par des techniques de clustering sur ces mêmes données utilisant des similarités plus classiques sur le contexte et les entités (à l'exception de l'approche par sacs de vecteurs qui ne peut pas s'appliquer à la classification au niveau de l'occurrence).

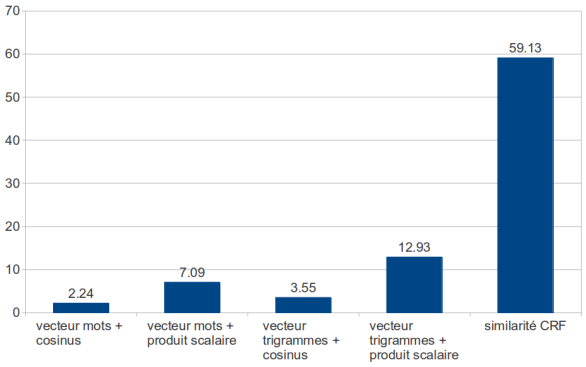


FIGURE 3: Évaluation des clusterings par rapport à la vérité terrain (ARI %)

L'intérêt de notre approche apparaît clairement. La prise en compte de la séquentialité est un élément important ; les résultats avec les n-grammes sont en effet meilleurs que des mots isolés, et ceux des CRF, qui prennent plus naturellement en compte cet aspect séquentiel, sont encore meilleurs. Les clusters obtenus par notre approche ne sont cependant pas exactement identiques à ceux de la vérité terrain.

Une analyse détaillée montre en effet qu'un cluster en particulier fait chuter les résultats en groupant des entités appartenant à deux classes distinctes de la vérité terrain. Ces classes qui semblent difficiles à distinguer sont celles du temps et des quantités. En effet, en l'absence d'informations autres que la forme des mots et les parties-du-discours, il semble impossible de distinguer des entités telles que '`sur les quatre derniers jours`' et '`sur les quinze derniers kilomètres`'.

6 Discussion, conclusion et perspectives

La résolution de problèmes d'apprentissage factices par les CRF permet de faire émerger des similarités au sein des séquences. Cette similarité tire ainsi parti de la richesse de description que permet les CRF (typiquement les parties-du-discours), ainsi que de la prise en compte naturelle de la séquentialité. On définit ainsi une similarité dans un espace non-métrique se voulant robuste grâce aux choix aléatoires répétés dans le processus. Bien sûr, ce principe est transposable à d'autres méthodes d'apprentissage, notamment les méthodes séquentielles stochastiques (HMM, MaxEnt...) ; l'utilisation des CRF, plus performants en général, est cependant plus naturelle.

Les évaluations menées sur deux tâches d'extraction d'informations mettent en valeur l'intérêt de l'approche, même si nous sommes bien conscients de la limite de l'évaluation d'une tâche de découverte qui oblige à la constitution d'une vérité terrain que l'on souhaite justement éviter. Enfin, il convient de préciser qu'il n'y a pas d'apprentissage sans biais, même pour l'apprentissage non supervisé (Mitchell, 1990). Ces biais représentent la connaissance de l'utilisateur et permettent de définir son problème. L'apport de connaissances sur les entités intéressantes, la description des séquences et des fonctions *features* sont autant d'informations permettant à l'utilisateur de canaliser la tâche de découverte sur son objet d'étude.

Plusieurs améliorations et perspectives sont envisageables à la suite de ce travail. D'un point de vue technique, l'étape de transformation des co-étiquetages en similarités, qui se contente dans nos expériences d'une simple normalisation, pourrait être approfondie. Il doit ainsi être possible d'utiliser d'autres fonctions (par exemple celles utilisées pour repérer des associations, expressions multi-mots, ou termes complexes complexes : information mutuelle, Jaccard, log-vraisemblance, χ^2...) pour obtenir des similarités encore plus fiables. Cela permettrait de pallier la faible robustesse de notre algorithme de clustering qui peut fusionner deux clusters sur le simple fait de quelques entités fortement connectées avec beaucoup d'autres nœuds. Des variantes sur l'étape de clustering peuvent aussi être envisagées. Il est par exemple possible d'utiliser des algorithmes de clustering hiérarchique. Il est aussi possible d'utiliser directement les similarités pour d'autres tâches, comme la recherche d'informations, le lissage pour des modèles de langues... D'un point de vue pratique, il serait intéressant d'obtenir une définition explicite de la similarité en récupérant les λ_i et μ_i avec les fonctions f et g associées. Cela permettrait d'appliquer la fonction de similarité à de nouveaux textes sans refaire les coûteuses étapes d'apprentissage, mais cela nécessite d'être capable de combiner les différentes fonctions de décodage utilisées pour l'application des différents modèles.

Références

COLLINS, M. et SINGER, Y. (1999). Unsupervised models for named entity classification. *In Proceedings of Empirical Methods for Natural Language Processing (EMNLP) conference*.

CONSTANT, M., TELLIER, I., DUCHIER, D., DUPONT, Y., SIGOGNE, A. et BILLOT, S. (2011). Intégrer des connaissances liguistiques dans un CRF : Application à l'apprentissage d'un segmenteur-étiqueteur du français. *In Actes de Traitement Automatique du Langage Naturel, TALN'11*, Montpellier, France.

EBADAT, A. R., CLAVEAU, V. et SÉBILLOT, P. (2012). Semantic clustering using bag-of-bag-of-features. *In Actes de le 9e conférence en recherche d'information et applications, CORIA 2012*, Bordeaux, France.

ELSNER, M., CHARNIAK, E. et JOHNSON, M. (2009). Structured generative models for unsupervised named-entity clustering. *In Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT-NAACL 2009)*, Boulder, Colorado.

FORT, K. et CLAVEAU, V. (2012). Annotating football matches : influence of the source medium on manual annotation. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

GOLDWATER, S. et GRIFFITHS, T. L. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. *In Proceedings of the ACL*.

GRAVIER, G., BONASTRE, J.-F., GEOFFROIS, E., GALLIANO, S., TAIT, K. M. et CHOUKRI, K. (2005). ESTER, une campagne d'évaluation des systèmes d'indexation automatique. *In Actes des Journées d'Étude sur la Parole, JEP, Atelier ESTER2*.

HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. New York : Springer.

HUBERT, L. et ARABIE, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

KAZAMA, J. et TORISAWA, K. (2007). Exploiting wikipedia as external knowledge for named entity recognition. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, Prague. Association for Computational Linguistics.

KOZAREVA, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. *In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics : Student Research Workshop*, pages 15–21, Trento, Italy.

LAFFERTY, J., MCCALLUM, A. et PEREIRA, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In International Conference on Machine Learning (ICML)*.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

LIU, B., XIA, Y. et YU, P. S. (2000). Clustering through decision tree construction. *In Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, pages 20–29, New York, NY, USA. ACM.

MANNING, C., RAGHAVAN, P. et SCHÜTZE, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

MERIALDO, B. (1994). Tagging english text with a probabilistic model. *Computational Linguistics*, 20:155–171.

MITCHELL, T. M. (1990). The need for biases in learning generalizations. *Rutgers Computer Science Department Technical Report CBM-TR-117, May, 1980. Reprinted in Readings in Machine Learning*.

NGUYEN XUAN VINH, Julien Epps, J. B. (2010). Information theoretic measures for clusterings comparison. *Journal of Machine Learning Research*.

PRANJAL, A., DELIP, R. et BALARAMAN, R. (2006). Part of speech tagging and chunking with HMM and CRF. *In Proceedings of NLP Association of India (NLPAI) Machine Learning Contest*.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850.

RAVI, S. et KNIGHT, K. (2009). Minimized models for unsupervised part-of-speech tagging. *In Proceedings of ACL-IJCNLP 2009*, pages 504–512.

RAYMOND, C. et FAYOLLE, J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. *In Actes de Traitement Automatique des Langues Naturelles, TALN'10*, Montréal, Canada.

RICHARD, D. et BENOIT, F. (2010). Semi-supervised part-of-speech tagging in speech applications. *In Interspeech 2010*, Makuhari (Japan).

SCHRAUDOLPH, N. N., YU, J. et GÜNTER, S. (2007). A stochastic quasi-Newton method for online convex optimization. *In Proceedings of 11th International Conference on Artificial Intelligence and Statistics*, volume 2 de *Workshop and Conference Proceedings*, pages 436–443, San Juan, Puerto Rico.

SHI, T. et HORVATH, S. (2005). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138.

SMITH, N. et EISNER, J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. *In Proceedings of ACL*.

van DONGEN, S. (2000). *Graph Clustering by Flow Simulation*. Thèse de doctorat, Université d'Utrecht.

WANG, T., LI, J., DIAO, Q., WEI HU, Y. Z. et DULONG, C. (2006). Semantic event detection using conditional random fields. *In IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06)*.

WANG, W., BESANÇON, R., FERRET, O. et GRAU, B. (2011). Filtering and clustering relations for unsupervised information extraction in open domain. *In Proceedings of the 20th ACM international Conference on Information and Knowledge Management (CIKM)*, pages 1405–1414, Glasgow, Scotland, UK.

WANG, W., BESANÇON, R., FERRET, O. et GRAU, B. (2012). Evaluation of unsupervised information extraction. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie.

WENHUI LIAO, S. V. (2009). A simple semi-supervised algorithm for named entity recognition. *In Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*, pages 58–65, Boulder, Colorado, USA. Association for Computational Linguistics.

