Apprentissage actif pour l'annotation de documents

Loïc Lecerf — Boris Chidlovskii

Xerox Research Centre Europe
6, chemin de Maupertuis,
F–38240 Meylan
{prenom.nom}@xrce.xerox.com

RÉSUMÉ. *Dans le cadre du projet LegDoc au Centre Européen de Recherche de Xerox, nous avons développé des composants pour l'annotation sémantique de documents semi-structurés. Alors que certaines entités sémantiques ont une forme régulière et peuvent être facilement extraites, d'autres collections plus complexes et hétérogènes nous ont amenés à déployer des méthodes d'apprentissage automatique. Dans les cas réels nous sommes souvent confrontés au problème technique de la non disponibilité de corpus annotés, pour des tâches d'annotations spécifiques. Comme l'annotation manuelle est coûteuse et propice à l'erreur, notre approche consiste à appliquer des méthodes d'apprentissage actif afin de considérablement réduire le corpus nécessaire à l'élaboration d'un modèle pertinent. Dans cet article, nous expliquons comment le principe de l'apprentissage actif est adapté à l'annotation interactive de documents orientés mise en page. Pour une utilisation efficace de l'apprentissage actif sur les grandes collections, nous déployons un classifieur probabiliste basé sur le principe de l'entropie maximum ainsi que trois métriques d'incertitude. Nous présentons le prototype ALDAI (Active Learning Document Annotation) et décrivons ses fonctionnalités ainsi que les choix d'implémentation. Le prototype offre une interface WYSIWYG, un haut langage pour la définition des attributs et intègre le composant d'apprentissage actif qui vise à assister l'utilisateur dans le processus d'annotation. Nous rapportons aussi des résultats de tests d'évaluation des techniques d'apprentissage actif sur une collection de données publiques (UCI) et une collection de documents internes.*

ABSTRACT. *In the framework of the LegDoc project at Xerox Research Centre Europe, we are developing components for the semantic annotation of semi-structured documents. While certain semantic entities have regular forms and might be easily extracted, more complex and heterogeneous collections favor the deployment of machine learning methods. Moreover, real world cases pose the technical challenge of the unavailable training sets for specific annotation tasks. As the manual annotation is costly and error-prone, our approach consists in applying active*

CORIA 2007, pp. 93–107
Saint-Étienne, 28–30 mars 2007

learning methods in order to considerably reduce the corpus required for accurate learning models. In this paper, we explain how the active learning principles get adapted the interactive semantic annotation of layout-oriented documents. We deploy the maximum entropy classifier for the probabilistic reasoning and three uncertainty metrics for the efficient application of active learning on large collections. We present the Active Learning Document Annotation Interface (ALDAI) prototype and describe its functionality and implementation choices. The prototype offers a WYSIWYG interface, a high-level language for feature definitons and integrates the active learning component aimed at helping users during the annotation process. We also report some evaluation results of testing the active learning techniques on one public (UCI) and one internal document collections.

MOTS-CLÉS : *Annotation sémantique, apprentissage actif, XML*

KEYWORDS: *Semantic annotation, active learning, XML*

1. Introduction

Les grandes compagnies et organisations possédant de grands fonds documentaires font face à de nombreuses difficultés pour faire migrer leurs documents vers un nouveau format qui leur permettrait de les déployer et de les réutiliser de façon plus efficace. Une telle efficacité peut être obtenue grâce à l'extraction de méta-données et d'informations dans le document. Le nouveau formalisme standard pour l'encodage de méta-informations et de données d'échange est le XML. Ce dernier, recommandé par le W3C, permet de définir un vocabulaire et une syntaxe adaptés aux données et facilite leur échange et la réutilisation du contenu. Les technologies construites autour de lui offrent de nouvelles fonctionnalités ; les recherches peuvent par exemple devenir plus significatives grâce à un balisage sémantique très précis sur les parties importantes du document. La conversion de document n'a donc pas seulement pour objectif de convertir un document d'un format en un autre, mais elle doit aussi permettre d'annoter de façon explicite les informations implicites encodées dans le format d'origine.

Au Centre Européen de Recherche de Xerox, le projet LegDoc (*Legacy Document Conversion*) a été entrepris, lequel a pour objectif d'automatiser différentes sous-tâches constitutives de la conversion d'un document en XML (Chanod *et al.,* 2005). Un cas typique de conversion commmence avec des documents disponibles en PDF, Postscript ou Microsoft Word, et un schéma pour les document cibles, défini sous la forme d'une DTD ou d'un XML Schema. Nous distinguons trois niveaux d'annotation. Les annotations relatives à la **mise en page** en terme de rendu physique des élements, telles que leurs positions x et y, leur largeur, hauteur, etc. Un second niveau, plus abstrait, fait référence à la **structure logique** du document et exprime les relations spatiales entre les élements dans une page telle que les colonnes, en-têtes, paragraphes et lignes. Le troisième niveau d'annotation est **sémantique** et fait réference à la signification de l'élement plus qu'à son apparence sur la page.

Les collections de documents peuvent varier aussi bien dans le contenu (document technique, livre, facture, catalogue, manuel etc.) que dans la forme (document papier ou électronique). La conversion de fonds documentaires en XML sémantiquement dense est habituellement effectuée par un expert du domaine et donc manuelle et très coûteuse. Dans le projet LegDoc, des premiers travaux ont permis de développer des méthodes pertinentes pour automatiser l'annotation sémantique de document (Chanod *et al.,* 2005). Elles s'appuient sur des techniques d'apprentissage automatique qui permettent d'inférer des régles d'annotation à partir d'un corpus annoté. Malheureusement, dans de nombreux cas, nous ne possédons pas de tels corpus. En effet, chaque tâche d'annotation est différente et exige des données d'apprentissage spécifiques.

Dans cette article, nous proposons une nouvelle approche interactive pour l'annotation sémantique de documents basés sur des méthodes d'apprentissage actif. L'apprentissage actif permet de créer un modèle pertinent avec des données d'apprentissage peu nombreuses car bien choisies. Cette approche permet l'élaboration d'un corpus annoté à faible coût.

Nous allons premièrement présenter le problème de l'annotation sémantique. Nous verrons l'approche utilisée dans le cadre de l'apprentissage supervisé avec un corpus d'apprentissage. Ensuite, nous montrerons l'approche par apprentissage actif, le principe, un bref état de l'art, ainsi que les choix techniques effectués. Enfin, dans une dernière partie, nous présenterons ALDAI (*Active Learning Documents Annontation Interface*), une interface pour l'annotation interactive de document. Ce système s'intègre au projet LegDoc. Il assiste l'utilisateur dans l'annotation de documents grâce à la combinaison d'une interface d'annotation et de navigation au sein du document et d'un composant d'apprentissage.

2. Le problème de l'annotation sémantique

La chaîne de traitement de la conversion documentaire est séquentielle. Chaque composant de la chaîne permet d'améliorer la qualité du document en cours de transformation en se rapprochant incrémentalement de son annotation finale. Les annotations de type mise en page ou structurel issues des étapes précédentes vont nous permettre d'inférer des informations précieuses pour notre phase d'annotation sémantique.

L'annotation sémantique peut être de différentes granularités. Deux types familliers sont les métadonnées et les entités. Les métadonnées réfèrent aux élements qui décrivent le document entier comme le titre, l'auteur, la date de création, etc. ; tous ces éléments sont habituellement indexés et utilisés pour les applications de gestion du contenu. Les entités sont des élements du contenu du document de faible granularité tels que les noms de personne et de société.

La Figure 1 offre un aperçu de l'interface ALDAI dont on discutera dans la suite de l'article. Elle permet de donner une idée de la tâche d'annotation. Ici une publication de recherche est en train d'être annotée en mode WYSIWYG. Le document est de structure arborescente avec au moins trois niveaux référencés comme *page*, *paragraphe* et *ligne*. Sur cette image, les classes sémantiques assignées aux annotations sont montrées par des couleurs différentes. Ici, l'utilisateur souhaite annoter par **title**, **author**, **affiliation**, **e-mail**, **reference**,ou **ignore** les éléments du document.

Certaines entités sémantiques, comme une addresse e-mail ou une date, ont des formes régulières ; écrire des expressions régulières à la main peut être suffisant pour capturer de telles entités dans un document. Dans des cas plus complexes, des méthodes d'apprentissage automatique peuvent être déployées pour inférer des régles d'annotation. Le principe de l'apprentissage supervisé requiert de construire un modèle probabiliste et d'apprendre les paramètres du modèle sur les corpus disponibles.

Annoter un document soulève certains problèmes. En effet, les annotations peuvent varier considérablement en taille, d'un important fragment comme une section de plusieurs pages, à un mot ou une phrase. Elle peuvent varier aussi en structure. Ainsi, un fragment de document à annoter peut correspondre à une feuille d'un arbre XML ou un sous-arbre.

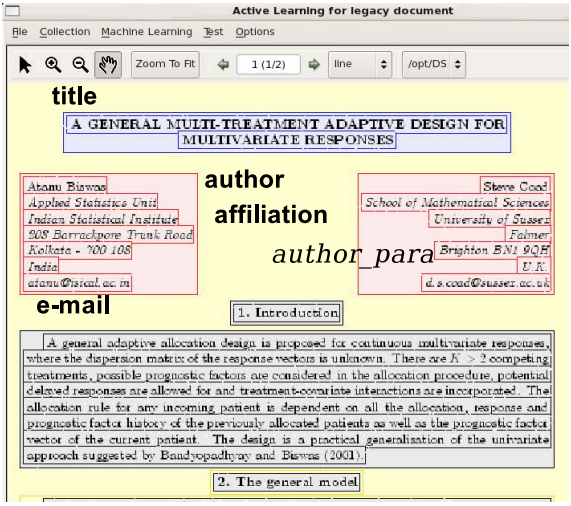

Figure 1. *Fenêtre d'annotation de document.*

3. Apprentissage automatique

L'approche de base pour différentes statégies d'annotation est l'utilisation de méthodes de classification probabiliste supervisées. Un classifieur probabiliste crée un modèle d'apprentissage, cohérent avec les exemples annotés par un expert, et qui est le plus général possible, c'est-à-dire performant sur des données non vues en apprentissage. Le but recherché est de déterminer des valeurs de caractéristiques propres à chaque classe pour déterminer, à partir d'un x, les probabilités de chacune des classes possibles. Formellement, un classifieur probabiliste cherche à estimer la probabilité conditionnelle $p(y|x)$ qui définit la probabilité que la classe de l'observation x soit y. Si le classifieur probabiliste est utilisé seul, la classe estimée est la classe pour laquelle cette probabilité est maximale. Dans le cadre de l'annotation, chaque x est un vecteur de caracteristiques d'une entité du document et chaque y est une annotation possible à associer avec les entités.

Pour être utilisées avec des méthodes d'apprentissage existantes, les instances à classer (les observations x) doivent être projetées dans un modèle des données consistant. Dans ce modèle, une instance va être vue comme un vecteur de caractéristiques permettant de décrire le plus fidèlement possible les spécificités des instances d'une même classe. Nous classons ces caractéristiques suivants trois catégories différentes qui récupèrent des informations utiles pour la discrimination.

3.1. *Extraction des attributs*

Une définition soignée des attributs $f(x, y)$ est primordiale pour de nombreux problèmes d'annotation. Dans une étape préliminaire, nous allons chercher à extraire, à partir du document XML, les attributs qui nous seront utiles pour l'apprentissage.

3.1.1. *Les attributs de contenu*

La première source d'informations que nous pouvons utiliser concerne les fragments textuels du document. Ces attributs permettent de décrire précisement les caractéristiques spécifiques aux chaînes de caractères. Nous pouvons penser, entre autres, aux nombres de caractères de la chaîne, à la présence de caractères spéciaux ou encore au caractère numérique de la chaîne.

Par exemple, on peut définir l'attribut f_1 tels que $f_1(x, y) = 1$ si y=**title** et le texte dans x est composé uniquement de caractères en majuscule ; 0 sinon.

3.1.2. *Les attributs de structure*

La deuxième source d'attributs qui est à notre disposition concerne la structure de l'arbre XML. Il est en effet parfois pertinent de connaître le contexte d'une feuille en explorant les noeuds autour.

On peut définir ici un attribut $f_2(x, y)$=1 si y=**affiliation** et le nom du noeud père est `paragraphe` ; 0 sinon.

3.1.3. *Les attributs de contenu XML*

Enfin, la dernière source d'attributs concerne les valeurs des attributs des noeuds XML dans l'arbre source.

Par exemple $f_3(x, y)$=1 si y=**title** et la valeur de l'attribut XML font de x est *times*, 0 sinon.

En utilisant cette représentation, un noeud peut être projeté dans ce modèle et le classifieur probabiliste travaille alors sur cette représentation simplifiée. Les attributs extraits dont nous disposons sont de types hétérogènes (discrets, numériques ou booléens) mais principalement à valeurs discrètes.

3.2. *Principe du maximum d'entropie*

Notre classifieur actuel est basé sur le principe du maximum d'entropie (Berger *et al.,* 1996) communément appelé MaxEnt.

Avec les contraintes basées sur les attributs choisis $f_j(x, y)$, la méthode du maximum d'entropie va chercher à maximiser la probabilité conditionnelle de $f_j(x, y)$. Il fait l'hypothèse qu'elle suit une loi exponentielle :

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_j \lambda_j \cdot f_j(x,y)\right), \quad [1]$$

où $Z(x)$ est un facteur de normalisation qui permet d'assurer que la valeur obtenue est une probabilité.

$$Z(x) = \sum_y exp\left(\sum_j \lambda_j \cdot f_j(x,y)\right). \quad [2]$$

Les valeurs λ_j représentent une pondération des attributs et permettent de déterminer un modèle pour lequel la distribution définie soit la plus exacte possible pour les données de l'ensemble d'apprentissage. Pour chaque choix de $\lambda = (\lambda_{j1}, \lambda_{j2}, ..., \lambda_{jm})$ que nous pouvons faire, nous définissons donc un modèle différent, le classifieur MaxEnt va déterminer parmi toutes ces possibilités le modèle optimal, en utilisant le principe de maximum d'entropie. Ce principe privilégie les modèles les plus uniformes et permet de trouver un maximum local. Pour l'estimation itérative des paramètres du modèle, nous utilisons la méthode quasi Newton, appelée aussi LBFGS qui est plus efficace que les méthodes habituelles GIS et IIS (Malouf, 2002).

4. Apprentissage actif

Bien que l'apprentissage du modèle et le choix des attributs soient très importants, le principal problème pour l'utilisation des méthodes d'apprentissage réside dans la non disponibilité de corpus annotés pour des tâches spécifiques d'annotations. L'annotation manuelle est difficile et très coûteuse. C'est pourquoi nous avons besoin d'un composant d'apprentissage capable d'aider l'utilisateur dans l'élaboration des données d'apprentissage.

La notion d'apprentissage actif fait référence à une méthode où l'algorithme d'apprentissage sélectionne à chaque itération les instances à annoter et les inclut dans la série d'exemples d'apprentissage. Ceci permet très souvent de réduire de façon importante la quantité de données nécessaires pour apprendre un modèle d'apprentissage supervisé. Au lieu d'annoter des instances aléatoirement pour obtenir les données d'apprentissage, le module d'apprentissage actif suggère d'annoter les instances dont il espère que le bénéfice sur l'apprentissage soit maximum. Le choix de l'élément à annoter est donc primordial. Les deux approches les plus répandues sont les méthodes basées sur l'incertitude et les méthodes basées sur les comités.

Dans les méthodes basées sur les comités, un ensemble de classifieur est créé. Les classifieurs sont entrainés avec les éléments annotés puis appliqués sur les éléments non annotés. L'élément choisi pour la prochaine annotation est celui dont le désaccord de prédiction entre les classifieurs est le plus grand. Dans (Seung *et al.,* 1992), les

auteurs ont introduit cette approche qu'il ont appellé *query by committee*. Différentes méthodes basées sur les comités sont apparues ces dernières années. *Decorate* en est une qui cherche à créer des classifieurs les plus différents possibles (Melville *et al.*, 2004).

Une autre approche consiste à utiliser une mesure d'incertitude (Lewis *et al.*, 1994b). Le plus souvent appelée *uncertainty based sampling*, cette méthode va chercher à proposer à l'annotateur d'annoter l'élément le moins certain ou le moins confiant pour le classifieur courant. A chaque itération, les éléments annotés forme le classifieur courant. Ce dernier permet de définir une mesure de confiance pour chaque élément non annoté. Cette méthode est simple d'utilisation. Elle permet en outre d'éviter la gestion de plusieurs classifieurs.

Ces deux approches ont eu des applications diverses dans la catégorisation de texte, le traitement automatique de la langue, la classification d'images etc. (Thompson *et al.,* 1999, Lewis *et al.,* 1994a).

Error rate reducing (Roy *et al.,* 2001) est une technique alternative. Cette technique consiste à sélectionner l'exemple selon un critère optimal : réduire le taux d'erreur sur les futures données test. Pour cela, il faut évaluer les données test après avoir simuler l'ajout de chaque élément non annoté, associé avec chacune des classes possibles. L'approche est pertinente mais reste difficilement utilisable pour les données de taille importante.

Enfin, une dernière approche consiste à combiner plusieurs algorithmes d'apprentissage actif. L'idée ici est de pouvoir basculer d'un algorithme d'apprentissage actif à un autre (Baram *et al.,* 2004, Osugi *et al.,* 2005). On peut par exemple comme le propose (Osugi *et al.,* 2005) alterner une *Exploration* des données non annotées et une *Exploitation* des données déja annotées. L'exploration se fait en annotant l'élément le plus différent des données annotées et l'exploitation consiste à annoter l'élément qui aura le meilleur bénéfice sur le classifieur.

4.1. *Mesure de l'incertitude*

Comme nous avons vu précédement en section 3.2, nous avons fait le choix technologique d'utiliser le classifieur MaxEnt. MaxEnt est en effet un classifieur flexible, qui peut facilement combiner des attributs de type syntaxique, sémantique ou pragmatique. Il a de plus deja montré son efficacité dans les domaines liés au traitement de la langue (Berger *et al.,* 1996). Il nous a semblé naturel, dans le cadre de notre projet, d'aborder l'apprentissage actif avec les méthodes basées sur l'incertitude et MaxEnt qui est un classifieur probabiliste.

L'incertitude d'un classifieur pour une prédiction donnée peut être évaluée de différentes façons. (Tong, 2001) cite trois mesures pour évaluer la confiance du module d'apprentissage sur ses prédictions. Soit une distribution des probabilités $P(y|x)$, pour une observation x, nous considérons les métriques suivantes :

1) Métrique différence :

$$conf\_diff(x) = p_1 - p_2 \quad [3]$$

avec p_1 et p_2 la première et la deuxième probabilité la plus importante dans la distribution de $P(y|x)$.

2) Métrique produit :

$$conf\_prod(x) = \prod_y P(y|x) \quad [4]$$

3) Métrique maximum d'entropie :

$$conf\_ME(x) = \sum_y P(y|x) log(P(y|x)) \quad [5]$$

Les valeurs de confiance calculées par ces métriques sont interprétées ainsi :

– La métrique "différence" est toujours positive. Une valeur faible indique une grande incertitude.

– La métrique "produit" est aussi toujours positive, néanmoins, ici une valeur faible correspond à une confiance élevée. En effet, nous avons une valeur maximale lorsque toutes les probabilités de la distribution sont égales.

– La métrique "maximum d'entropie" est négative. Ses valeurs se rapprochent de 0 lorsque le classifieur devient certain.

5. Evaluation

Nous présentons une série de tests effectués sur une collection client composée de documents techniques contenant chacun plusieurs centaines de pages, ainsi que sur une série de données libres disponibles sur le site de UCI (D.J. Newman *et al.*, 1998). Cette dernière collection est composée de données de petite taille mais intéressantes par leur diversité en terme de nombre de classes ou d'attributs. Nous avons simulé l'apprentissage actif, en annotant à chaque itération les n éléments les plus incertains pour le classifieur selon les métriques d'incertitude définies précédemment (Avec n petit au départ et augmentant incrémentalement). Après chaque itération, le modèle est réappris avec l'ensemble des éléments annotés et sa qualité est évaluée. Pour chaque évaluation, nous effectuons une validation croisée avec 4/5 des données formant l'ensemble d'apprentissage et 1/5 les données tests. Les éléments de l'ensemble d'apprentissage, sont tour à tour annotés, et à chaque itération le classifieur est appris sur ces données annotées puis testées sur l'ensemble de test. Ceci nous

permet de voir l'évolution de la qualité du modèle tout au long du processus d'apprentissage actif. Les trois métriques d'incertitude sont comparées avec la métrique aléatoire (*Random*) où les éléments à annoter sont choisis au hasard. Dans la collection de documents techniques, nous cherchons à capturer les titres de section. Nous utilisons ici la mesure *F1* qui permet de combiner la précision et le rappel. Pour la collection UCI, l'évaluation est effectuée par la mesure d'exactitude qui est le nombre d'éléments correctement annotés sur le nombre total d'éléments.

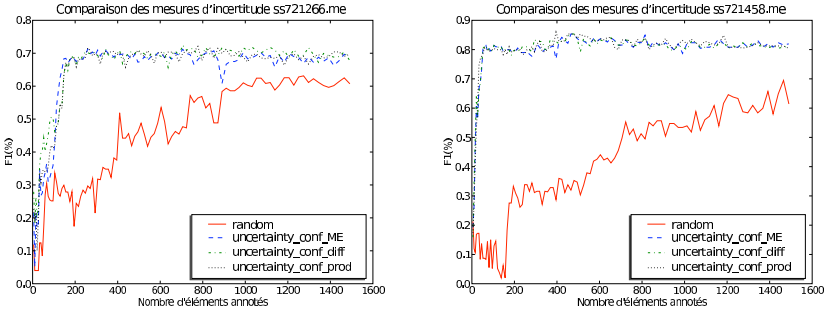

Figure 2. *Evaluation sur une série de documents techniques*

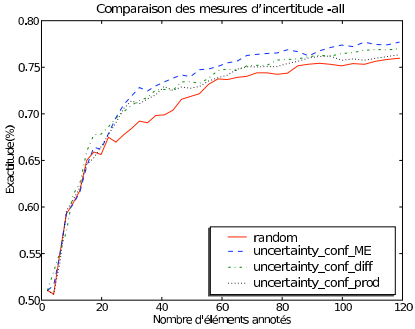

Figure 3. *Moyenne des évaluations sur 10 documents d'UCI.*

La figure 2 présentent l'évaluation des différentes mesures d'incertitude pour la collection de documents techniques. Nous montrons deux exemples typiques de la collection. Les documents ss721266 et ss721458 sont composés respectivement de 12781 et 7030 lignes. L'apprentissage actif est simulé jusqu'à l'annotation de 1500 élements, soit environs 11% et 21% du document. Sur la collection UCI, la pertinence des résultats est moins évidente car les données sont pour la plupart de petite taille.

La figure 3 présente une moyenne des résultats pour l'annotation interactive de 10 documents contenant entre 150 et 3190 instances. L'apprentissage actif est simulé pour l'annotation des 120 premières instances. Ceci nous donne une tendance générale de l'évolution de la qualité du modèle. Nous avons pu voir que les trois mesures dépassent la mesure aléatoire. Nous avons aussi remarquer que l'utilisation de la mesure basée sur le principe de l'entropie maximum permet sur les tests effectués une légère amélioration par rapport aux deux autres métriques en particulier sur la collection UCI. Le gain réel apporté par l'apprentissage actif en terme de nombres d'éléments annotés est très variable d'une collection à l'autre, pouvant atteindre jusqu'à plus de 95% comme on peut le voir sur les documents issus de la première collection.

6. ALDAI : une interface pour l'annotation interactive

La difficulté technique et le coût de l'annotation nous ont amenés à développer ALDAI. ALDAI est une interface pour l'annotation interactive. Elle se situe à la fin de la chaîne de traitement du projet LegDoc et profite donc de l'ensemble des étapes en amont. L'objectif est donc d'assister l'utilisateur dans la création d'un corpus annoté nécessaire au développement d'un modèle pertinent, applicable au reste du document ou à une collection de documents.

Le système ALDAI est formé de 2 composants principaux : une interface utilisateur simple et intuitive pour la visualisation du document et l'annotation manuelle ainsi qu'un composant d'apprentissage comprenant un ensemble d'outils pour l'assister dans le processus d'annotation.

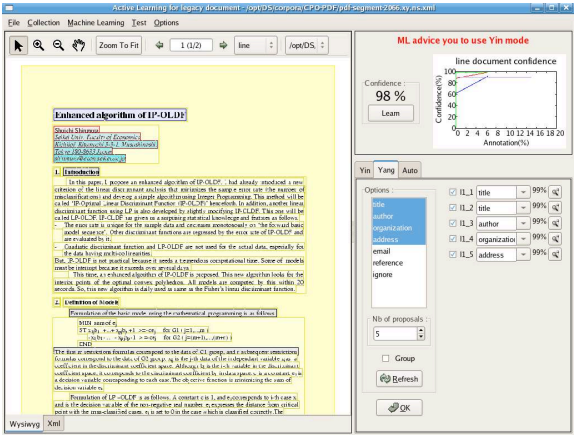

Figure 4. *Vue d'ensemble de ALDAI.*

6.1. *Composant interface*

L'interface travaille sur les documents XML issus de la conversion des documents PDF. Elle offre la possibilité d'avoir une vue structurée du document XML ou bien une vue WYSIWYG obtenue grace à une application sur le document d'une feuille de style approprié.

Au mode WYSIWYG est intégré une barre d'outils permettant de naviguer simplement à travers les documents d'une collection, les pages d'un document ou une zone particulière d'une page grâce à une fonction de zoom. De même, un système de cadres colorés permet de visualiser facilement les annotations semantiques. Ces cadres permettent aussi de mettre en valeur les éléments concernés par les propositions courantes du module d'apprentissage. Deux menus contextuels sont intégrés au document pour la visualisation des attributs extraits d'un élément et pour l'annotation manuelle.

Toutes les annotations sont gérées au sein du document. Elles ne changent pas la structure hiérarchique du document mais sont injectées sous la forme d'une paire $attribut = valeur$ dans le noeud correspondant. Ceci permet facilement de charger un document pour reprendre l'annotation, de sauver ou de charger un modèle appris lors d'une session précédente.

6.2. *Composant d'apprentissage actif*

L'idée principale sur laquelle repose ALDAI est de réunir un composant adaptable d'apprentissage au sein d'un environnement d'annotation de documents. Premièrement, grâce au développement de techniques d'apprentissage actif, le système développé permet de guider le processus d'annotation, par la sélection du prochain élément à annoter et donc de réduire la taille de l'échantillon d'apprentissage nécessaire et limiter le risque d'annotations erronées. Deuxièmement, elle peut estimer la précision et le coût pour l'annotation automatique du reste du document.

6.2.1. *Gestion des attributs*

Il est très important de pouvoir définir le mieux possible les attributs $f(x, y)$ à extraire. ALDAI propose un module avancé pour la gestion des attributs. Par défaut, il offre un ensemble d'attributs habituellement utilisé issus des trois sortes d'informations disponibles dans le document : les caractéristiques de contenu, de structure et celles issues des attributs XML. En plus, ALDAI offre la possibilité de définir de nouveaux attributs plus spécifiques ou d'en modifier d'autres, ceci grâce à un langage de haut niveau. Les attributs $f(x, y)$ associent une étiquette y à un élement x en utilisant une paire $(chemin, fonction)$ où chemin est une expression XPATH sur l'arbre XML qui cible un ou plusieurs noeuds et une fonction issue du lambda calcul Python.

Par exemple, l'attribut $[RBr\_font]$ est défini ainsi :

```
apply(lambda x :x.prop('font'),
 './following-sibling : :*[1]')
```

Elle donne la valeur de l'attribut police du noeud frère situé à droite du noeud courant x.

Dans les documents arborescents, les attributs peuvent être définis et groupés par le type de sous-arbres x. Plusieurs types de sous-arbres peuvent avoir des attributs spécifiques. Par exemple, un paragraphe peut avoir comme caractéristique l'alignement.

6.2.2. *Modes d'apprentissage*

Pour aider l'utilisateur dans le processus d'annotation, l'interface intègre deux modes d'apprentissage opposés, que l'on nommera par la suite le mode Yin et le mode Yang. Le mode Yin met plus l'accent sur l'amélioration du modèle que sur l'annotation. Dans ce mode, l'utilisateur investit dans le modèle en annotant les éléments les plus incertains pour le classifieur. Ceci va donc permettre d'augmenter la précision du modèle et de ses prédictions.

Dans le mode Yang, l'annotation prévaut sur l'apprentissage. Ce mode travaille avec les propositions les plus confiantes. Le coût associé est minimal car ces propositions sont souvent correctes et peuvent être validées par groupes (pour le coût d'un clic). Ceci permet de rapidement augmenter le nombre d'éléments annotés du corpus, mais ce mode améliore rarement la qualité du modèle.

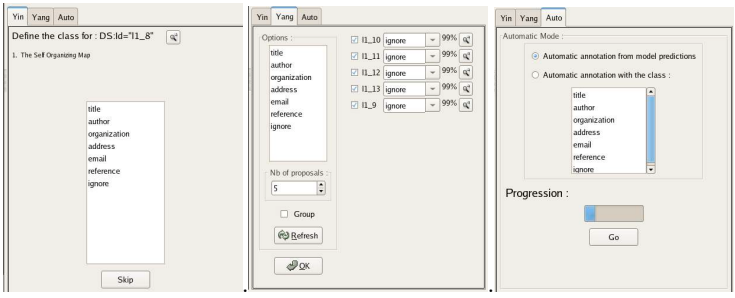

Figure 5. *Les 3 panneaux d'annotation de l'interface.*

Les deux modes sont complémentaires. Un long travail en mode Yin peut être très coûteux et même inutile à partir du moment où le modèle ne s'améliore plus. D'un autre côté, un travail trop long en mode Yang peut épuiser la capacité du modèle à faire des prédictions correctes. Il s'avère que la plus efficace statégie d'annotation est souvent un va-et-vient entre ces deux modes d'apprentissage. Notre interface intègre un algorithme d'estimation du coût qui suggère quand changer de mode. Une autre approche consiste à commencer en mode Yin puis passer en mode Yang dès que le modèle ne s'améliore plus.

Enfin, un mode automatique permet de stopper l'apprentissage et d'appliquer le modèle courant à tous les éléments non annotés du document ou d'une collection de document. L'utilisation d'une fenêtre spéciale permet de surveiller la confiance du modèle courant pendant tout le processus d'annotation. Cette confiance est la moyenne de l'incertitude pour chacun des éléments non annotés. Le passage en mode automatique reste à l'appréciation de l'utilisateur en fonction de cet indice de confiance.

La figure 5 montre les trois panneaux d'annotation proposés par le système. Dans le premier, l'annotateur travaille en mode Yin, ALDAI effectue un zoom sur un élément choisi par le composant d'apprentissage actif et lui propose de l'annoter avec une des classes possibles. Le second panneau correspond au mode Yang. Ici, l'utilisateur choisit les classes qu'il cherche à annoter, ALDAI fait un ensemble de propositions que l'annotateur pourra valider ou corriger. Enfin, le dernier panneau permet à l'utilisateur de lancer l'annotation automatique à partir des prédictions du modèle sur le reste du document.

7. Conclusion

Nous avons présenté nos travaux de recherche sur l'annotation de documents par apprentissage actif. Nous avons étudié différentes statégies d'annotation et évalué les méthodes. Ces travaux s'inscrivent dans le projet LegDoc qui a pour objectif de pouvoir convertir en masse des documents vers XML et d'offrir des techniques pour leur annotation sémantique. L'annotation sémantique est en effet essentielle pour le développement d'outils de recherche avancée. Nous avons aussi présenté ALDAI, une interface pour l'annotation interactive ; l'environnement d'annotation, les différents modes d'apprentissage ainsi que la gestion des attributs. Notre approche permet d'annoter sémantiquement des documents structurés, assistés par un module d'apprentissage.

Nous avons pu saisir l'intérêt offert par ces techniques. Nos travaux permettent de fortement réduire le coût de l'annotation grâce aux méthodes d'apprentissage actif. L'interface ALDAI qui intègre les méthodes développées se révèle être un outil précieux pour l'annotation de documents.

D'autres recherches sont en cours pour permettre d'intégrer des méthodes plus avancées. Par exemple, l'implémantation des *champs aléatoires conditionnels* permettrait de prendre en compte la structure des annotations du document et donc une meilleure précision (Lafferty *et al.,* 2001). De même pour l'apprentissage actif, nous cherchons à appliquer les méthodes récentes comme le *pré-clustering* (Nguyen *et al.,* 2004) qui permet de choisir d'annoter des éléments différents parmi les moins confiants ou les méthodes qui alternent *Exploration* et *Exploitation* du modéle (Osugi *et al.,* 2005). Le challenge consiste à trouver le moyen d'appliquer ces techniques sur les grandes collections tout en maintenant un système réactif.

8. Bibliographie

Baram Y., El-Yaniv R., Luz K., « Online Choice of Active Learning Algorithms », *J. Mach. Learn. Res.*, vol. 5, p. 255-291, 2004.

Berger A. L., Pietra S. D., Pietra V. J. D., « A Maximum Entropy Approach to Natural Language Processing », *Computational Linguistics*, vol. 22, n° 1, p. 39-71, 1996.

Chanod J.-P., Chidlovskii B., Dejean H., et al., « From Legacy Documents to XML : A Conversion Framework », *Proc. European Conf. Digital Libraries*, p. 92-103, 2005.

D.J. Newman S. Hettich C. B., Merz C., « UCI Repository of machine learning databases », 1998.

Lafferty J., McCallum A., Pereira F., « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data », *Proceedings of the 18th International Conference on Machine Learning (ICML)*, ACM Press, New York, NY, USA, 2001.

Lewis D. D., Catlett J., « Heterogeneous Uncertainty Sampling for Supervised Learning », *Proceedings of the 11th International Conference on Machine Learning(ICML)*, p. 144-156, 1994a.

Lewis D., Gale W., « A Sequential Algorithm for Training Text Classifiers », *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994b.

Malouf R., « A Comparison of Algorithms for Maximum Entropy Parameter Estimation », *Proceedings of the 6th Conference on Natural Language Learning*, p. 49-55, 2002.

Melville P., Mooney R. J., « Diverse Ensembles for Active Learning », *Proceedings of the 21th international Conference on Machine Learning (ICML)*, ACM Press, New York, NY, USA, p. 74, 2004.

Nguyen H., Smeulders A., « Active Learning Using Pre-clustering », *Proceedings of the 21th international Conference on Machine Learning (ICML)*, p. 79-86, 2004.

Osugi T., Kun D., Scott S., « Balancing Exploration and Exloitation : A New Algorithm for Active Machine Learning », *Proceedings of the 5th International Conference on Data Mining (ICDM)*, p. 330-337, 2005.

Roy N., Mccallum A., « Toward Optimal Active Learning through Sampling Estimation of Error Reduction », *Proceedings of the 18th International Conference on Machine Learning (ICML)*, p. 441-448, 2001.

Seung H. S., Opper M., Sompolinsky H., « Query by Committee », *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, p. 287-284, 1992.

Thompson C. A., Califf M. E., Mooney R. J., « Active Learning for Natural Language Parsing and Information Extraction », *Proceedings of the 16th International Conference on Machine Learning (ICML)*, p. 406-414, 1999.

Tong S., Active learning : theory and applications, PhD thesis, 2001. Adviser-Daphne Koller.