Paper presented at the: Journées Francophones d'Apprentissage (JFA'98), Arras 1998.
Available at http://www-poleia.lip6.fr/~zaragoza

Modèle Hiérarchique de Recherche et d'Extraction de l'Information Textuelle de Surface.

Hugo Zaragoza, Patrick Gallinari.

LIP6, Université Pierre et Marie Curie,
4, place Jussieu F-75252,
PARIS cedex 05 (F).
{Hugo.Zaragoza, Patrick.Gallinari}@lip6.fr

Nous présentons une application de techniques d'apprentissage numérique au domaine de l'extraction d'information textuelle. Ce n'est que très récemment que les techniques d'apprentissage ont commencé à être utilisées dans ce domaine. Le système proposé est hiérarchique et réalise successivement la localisation, l'extraction et l'étiquetage des mots pertinents à une requête prédéterminée, au sein d'un corpus d'articles. Chaque niveau de la hiérarchie correspond à un niveau structurel du corpus (document, section, paragraphe, ...), cela nous permet d'une part d'éliminer l'information non pertinente dès que possible, d'autre part d'améliorer la sélection d'information pertinente à chaque niveau. Au dernier niveau, l'analyse de séquences de mots est faite à l'aide de modèles de Markov Cachés.

MOTS-CLES : Extraction de l'Information, Recherche de l'Information, Modèles Stochastiques de Langage.

1. Introduction

Les domaines de l'Extraction et de la Recherche de l'Information textuelle (respectivement EI et RI) sont l'objet de recherches actives depuis plusieurs années dans la communauté Intelligence Artificielle. Ce n'est pourtant que récemment et avec l'apparition de grands corpus de données que l'on a ressenti la nécessité d'intégrer dans les systèmes existant, ou même de leur substituer, des techniques d'apprentissage automatique. Cette problématique connaît une expansion rapide avec le traitement de documents électroniques et les diverses applications liées à ce type d'information. Cela est dû, en partie, à la demande de systèmes génériques rapidement prototypables dans la communauté EI, alors que les systèmes linguistiques actuels sont en général complexes et fortement dédiés. D'autre part, en RI, de nombreuses applications nécessitent des modèles plus riches et plus complexes que ceux dont on dispose aujourd'hui. En RI, la

mise en oeuvre de techniques simples de classification a rapidement permis d'obtenir de bons résultats, la situation est très différentes en EI qui reste encore largement le domaine de l'analyse linguistique. Le traitement de grands corpus de données induit des besoins qui se situent à la frontière de ces deux domaines et l'on assiste actuellement au début du rapprochement des méthodes utilisées par les deux communautés, avec en plus une utilisation de l'apprentissage. C'est dans cette optique que se situe le travail que nous présentons. Nous nous intéressons à l'extraction d'informations de surface, i.e. d'informations qui ne demandent pas un traitement linguistique complexe pour être catégorisées. Notre but est de détecter et d'extraire dans des textes, des passages ou des séquences de mots, contenant des informations pertinentes concernant un ensemble de requêtes. Nous proposons un système hiérarchique basé sur des techniques d'apprentissage numérique qui résout successivement un problème de filtrage (problématique RI) et un problème de compréhension de messages (problématique EI). Ce système est testé sur un problème typique d'extraction d'information, il s'agit de la tâche *Scenario Templates* (patrons d'événements) de MUC6 (*Sixth Message Understanding Conference*) [3] qui est devenu une des références dans le domaine.

Dans la section 1, nous introduisons cette tâche, et nous présentons brièvement les domaines EI et RI. Nous décrivons ensuite les modèles utilisés pour les étapes RI (section 2), et EI (section 3), enfin, nous donnons des résultats préliminaires (section 4) concernant l'évaluation du système.

1.1 Base de données et tâche

MUC est une conférence américaine qui, depuis 1993, met en compétition des systèmes d'EI dans des conditions de développement et d'évaluation rigoureuses [3]. Cette compétition fait partie depuis 1996 du programme DARPA : *TIPSTER Text Research*. Un de ses intérêts principaux a été la définition de tâches et de bases de données spécifiques pour le développement et l'évaluation des systèmes d'analyse de texte. C'est la seule compétition de ce type dans le domaine EI, et elle est devenue la référence principale pour la compréhension de messages; son évolution constitue une véritable historique de l'évolution du domaine. Les techniques employées par la communauté MUC sont principalement issues de la linguistique computationelle : il s'agit pour la plupart de systèmes à base de règle qui codent les connaissances du domaine de la tâche couplés à des analyseurs syntaxiques.

Le corpus sur lequel nous avons réalisé des tests est constitué d'articles du *Wall Street Journal*. La tâche consiste à extraire d'une série d'articles les informations pertinentes au mouvement de postes dans les entreprises (nominations, changements, licenciements). Plus spécifiquement, le système doit détecter les zones de textes pertinentes et leur affecter une étiquette conceptuelle parmi un ensemble fini d'étiquettes : pour chacun des mouvements détecté, il s'agit de remplir un patron qui contient une description en plusieurs champs de l'événement (voir figure 1, gauche). Le corpus contient 200 articles et 400 patrons. Dans le travail présenté, nous avons traité une partie de cette tâche qui consiste à détecter pour chaque mouvement les informations du type «poste occupé» et «identité de la personne» (deux champs parmi une dizaine au total dans un patron). La même démarche peut être appliquée pour détecter les informations concernant la plupart des autres champs, cela reste encore à faire. Nous avons utilisé pour l'apprentissage une base de texte pré-étiquetée. D'autres applications typiques de la démarche proposée concernent le fil-

trage de messages électroniques, le résumé automatique, la recherche d'informations sur le WWW.

Une caractéristique de cette tâche d'extraction est que 50% des textes et plus de 80% des paragraphes de ces articles ne contiennent pas d'informations pertinentes pour la tâche. Il s'agit d'une situation typique de ce type de tâche et ce pourcentage pourra être encore supérieur dans des applications destinées par exemple à traiter des informations accessibles sur Internet.

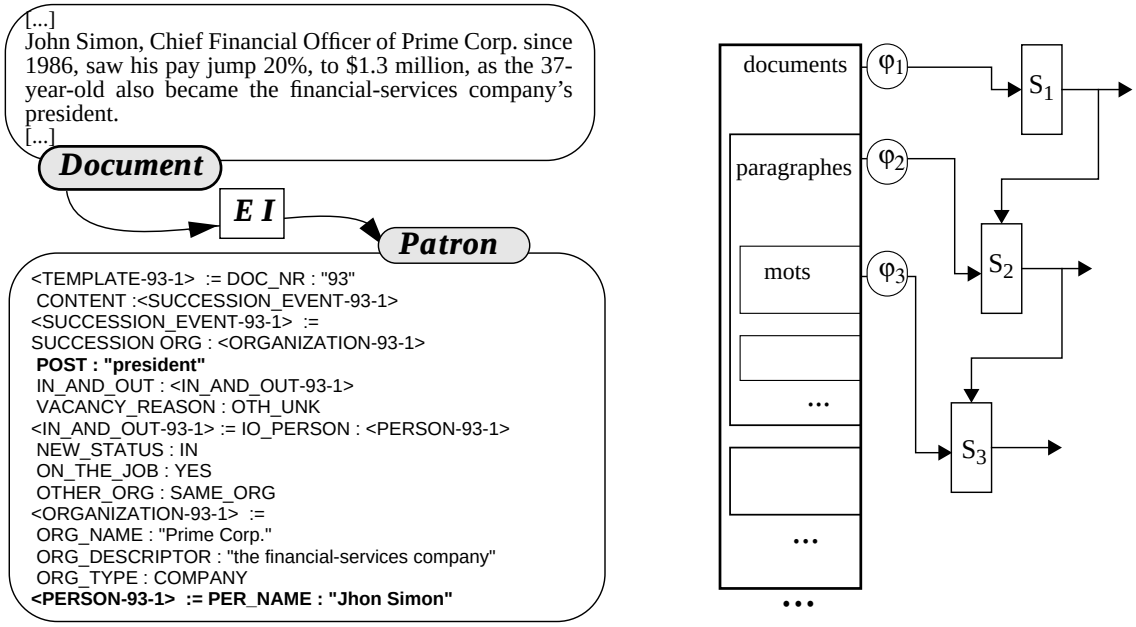

Figure 1: La tâche Scenario Templates de MUC6 (gauche) et le schéma du modèle d'Extraction Hiérarchique de l'Information proposé (droite). *Gauche: Le corpus est composé d'articles du Wall Street Journal (Document) et des patrons correspondants (Patron). En gras, les deux champs traités dans cet article. Droite: Les cercles (φ_i) représentent des différents codages et les carrés (S_i) des classifieurs. Trois niveaux structurels sont représentés : document, paragraphe et mot.*

L'utilisation directe de techniques EI pour extraire les informations pertinentes est coûteux, en effet, ces méthodes travaillent directement au niveau des séquences de mots. Etant donné les caractéristiques de la tâche, il parait intéressant d'éliminer le plus grand nombre possible de textes et passages non pertinents dans le corpus en travaillant au niveau d'unités globales, avant d'effectuer l'extraction proprement dite. Nous supposerons pour cela dans la suite qu'il est plus facile de localiser l'information que de l'extraire. Cette localisation peut être vue comme une analyse de type RI préalable à l'analyse EI, elle peut être faite au niveau du document, du paragraphe, et plus généralement sur plusieurs niveaux structurels successifs. L'approche hiérarchique que nous adoptons pour cette sélection nous permet, d'une part, d'éliminer dès que possible les informations non pertinentes, et d'autre part de raffiner successivement la sélection de variables (ou termes) nécessaires pour détecter l'information pertinente.

Après la localisation, il s'agit d'extraire les mots correspondant aux champs du patron. Nous faisons pour cela une analyse des séquences de mots dans les unités sélectionnées en utilisant une modélisation Markovienne. Le schéma de cette approche hiérarchique est représenté sur la figure 1 (droite).

1.2 Recherche et Extraction de l'Information

Le but initial de la RI était d'indexer et de retrouver rapidement, parmi de larges bases textuelles, des sous-ensembles (quelquefois ordonnés) de documents pertinents par rapport à une requête ou un sujet déterminé [13]. L'application typique de ces techniques est l'accès par requêtes aux bases de données de bibliothèques. Le champs de la RI s'est diversifié récemment et aborde actuellement des tâches plus spécifiques comme la catégorisation de messages électroniques ou la recherche interactive d'informations dans le WWW.

L'EI a pour but l'interprétation automatique du langage, et dans un cadre plus restreint la compréhension des messages simples contenus dans un texte. Ses applications sont l'interprétation des requêtes d'un utilisateur, la construction automatique de bases de données à partir de documents écrits, ou la construction de systèmes de dialogue [5]. L'EI a été utilisée principalement sur des domaines de connaissances limités. Les systèmes sont basés sur une approche linguistique, ils sont en général complexes et dédiés à des tâches très spécifiques. Ce n'est que très récemment que l'on a vu apparaître dans ce domaine des méthodes issues de l'Apprentissage. Cette nouvelle approche vise au développement de systèmes rapidement portables, capables de traiter de grosses masses de données, capables en plus de travailler avec des textes peu grammaticaux et bruités comme c'est le cas des retranscriptions d'interrogation en langage naturel ou des textes provenant des messageries électroniques.

A notre connaissance, il n'existe pas d'approches combinant la sélection hiérarchique dans la RI et des modèles stochastiques pour l'EI. Certains travaux récents ont traité de la sélection hiérarchique de variables pour la RI, comme, par exemple, Koller and Sahami [7], qui ont implémenté des arbres probabilistes à deux niveaux pour la catégorisation de textes, ou Wiene, Pedersen and Weigend [15] qui utilisent une approche similaire avec des réseaux de neurones. La combinaison de techniques d' EI et RI est très actuelle et il n'existe pas encore de travaux importants publiés; une approche symbolique a été proposé récemment par Robertson [10].

2. Localisation de Textes Pertinents

Nous donnons dans cette section une description des techniques de codage et de classification utilisées pour localiser des passages contenant de l'information pertinente, ces passages seront ensuite traités par des techniques d'EI. Par *passage* nous entendons une partie structurelle de texte (dans le cas de MUC6, des documents et/ou des paragraphes). Un passage est dit pertinent pour la tâche d'extraction s'il fait référence à au moins un patron concernant cette tâche.

Souvent les passages constituent une structure emboîtée, comme dans le cas des documents et des para-

graphes. Le traitement que nous réalisons est identique pour tous les passages. Avant la catégorisation, le texte est pre-traité. Nous utilisons pour cela les techniques classiques dans le domaine de la RI [13]: un anti-dictionnaire, l'élimination des mots de très haute et très basse fréquence, et l'algorithme Porter de *stemming*. Les passages sont ensuite codés comme des vecteurs, et classés selon leur pertinence. Le type de classifieurs utilisé est décrit dans la section 2.1 et la sélection de variables et le codage vectoriel des passages est décrit dans la section 2.2.

2.1 Classification à partir de Perceptrons Multi-Couches

Nous utilisons des Perceptrons Multi-Couches (PMCs) [4] pour la classification de passages. Ces systèmes nous fournissent une estimation de la probabilité a posteriori de la pertinence du passage. Cette mesure est utilisée pour ordonner tous les passages du corpus par ordre de pertinence et rejeter les moins pertinents. On peut aussi appliquer la théorie de la décision à partir de ces estimations, pour choisir un seuil optimal de rejet. Néanmoins, ce seuil dépendra du coût, choisi *a priori*, d'éliminer des documents pertinents et de conserver des documents non pertinents. Il est traditionnel en RI de présenter les performances des systèmes pour toutes les valeurs de ce seuil. Cela donne lieu a une courbe de performance, connue comme la courbe de précision-rappel (si n est le nombre de passages correctement classés comme pertinents, N le nombre total de passages pertinents, et N' le nombre total de passages classés comme pertinents, alors la précision est égal à n/N' et le rappel est égal à n/N).

Dans notre application, les documents non rejetés sont ensuite examinés plus en détail par les modules suivants. Pour cette raison, nous nous intéressons à des régimes de rappel élevé (peu de documents pertinents rejetés, au risque d'accepter des documents non pertinents). Notons que ce critère ne correspond pas au critère optimisé par l'algorithme d'apprentissage des réseaux de neurones.

2.2 Codage et Sélection Hiérarchique de Variables

Notre capacité à discriminer entre les passages pertinents et non pertinents dépend en grande partie des termes choisis pour représenter ces passages. La mesure utilisée ici pour la sélection des termes est la mesure dite U [1][6], elle est similaire à la mesure du χ^2 utilisée en statistique, mais ne prend pas en compte l'information négative (c'est à dire, l'absence de termes dans un passage). Si p et p' sont respectivement le nombre de passages pertinents et non pertinents, dans un corpus *B*, dans lesquels le mot w_i apparaît, q et q' le nombre de passages pertinents et non pertinents dans lesquels le mot n'apparaît pas, et N le nombre total de passages (N=p+p'+q+q'), la mesure U du mot w_i, u_i, est :

$$u_i = \varphi_U(w_i, B) = \sqrt{N} \cdot \frac{(pq' - p'q)}{\sqrt{(p + p') \cdot (p + q) \cdot (q + q') \cdot (p' + q')}}$$

Un passage sera représenté par ses T termes de plus grande mesure U. Cela peut être fait au niveau des paragraphes ou des documents. Dans la figure 2 nous présentons les 100 termes de plus grande mesure U

au niveau des paragraphes pour notre corpus. Les mots de plus grande mesure U sont tous proches, du point de vue de la sémantique de la tâche d'extraction. Le codage du passage réalisé à partir des termes sélectionnés dépend de la fréquence d'apparition de chaque terme dans des passages pertinents et non pertinents. La sélection au niveau paragraphe sera plus fine si on élimine, préalablement à la sélection, le plus grand nombre possible de documents non pertinents. Cela nous a mené vers une approche hiérarchique de la sélection de termes, de la façon suivante : d'abord, les termes sont sélectionnés au niveau des documents, les documents sont ensuite codés et classifiés et seuls les documents jugés pertinents sont gardés. Une nouvelle sélection de termes est alors effectuée, au niveau des paragraphes parmi les documents préalablement jugés pertinents. Pour le codage des passages nous utilisons la technique standard de codage *tf-idf* [13].

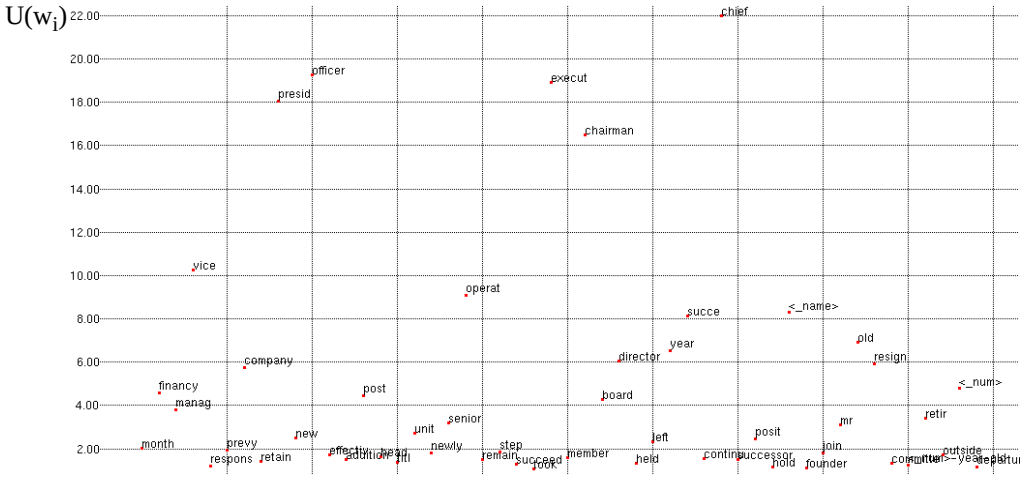

Figure 2: Mesure U des termes au niveau des paragraphes. L'axe vertical indique la mesure U des mots, la position horizontal est aléatoire (pour une meilleure visualisation).

Nous pouvons vérifier empiriquement l'intérêt de l'approche hiérarchique, en classant des paragraphes avec (SH) et sans (SNH) sélection hiérarchique des termes. Dans la figure 3 (gauche) nous montrons les courbes de précision-rappel pour les deux modèles: le modèle SH est supérieur au modèle SNH pour des valeurs élevées de rappel (ces valeurs correspondent aux régions qui nous intéressent).

Dans la figure 3 (droite) nous représentons la mesure U des termes par rapport aux modèles SH (axe horizontal) et SNH (axe vertical) précédents. Dans cette représentation, les termes qui apparaissent a gauche (SNH=2) sont éliminés par la SNH et les termes qui apparaissent en bas (SH<2) sont éliminés par la SH. Il est intéressant de noter le choix par SH de termes comme *company* ou *search*, clairement en relation avec notre tâche d'extraction (alors qu'ils sont éliminés par SNH), et l'élimination de termes comme *newspaper*, *automotor* ou *industry*, qui ne semblent pas intéressants pour cette tâche. Remarquons aussi que tous les termes éliminés par SH sont des termes avec des faibles mesures U, mais plusieurs termes introduits par la SH (comme ceux déjà mentionnés) ont une mesure U forte. Le reste des termes (84%) sont

sélectionnés par les deux modèles, et donc représentés dans la partie centrale de la figure.

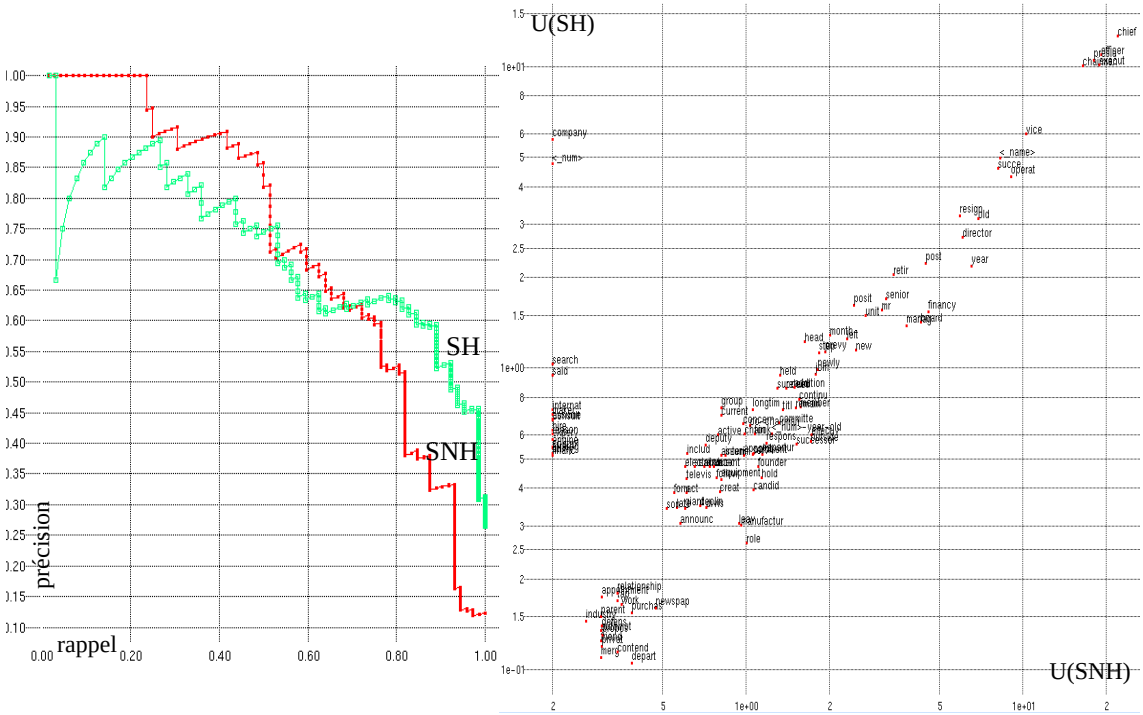

Figure 3: Sélection de Variables Hiérarchique. Courbes de précision-rappel (gauche) et mesure U des termes (droite) pour les modèles hiérarchique (HS) et non hiérarchiques (NHS).

3. Le Modèle d'Extraction Stochastique

L'approche statique des techniques RI (où les passages de texte sont codés comme des vecteurs de taille fixe) est difficile à étendre en EI. L'extraction d'information de type sémantique nécessite un traitement dynamique des mots dans un passage et la prise en compte du contexte.

Dans ce but, nous avons utilisé des modèles de Markov cachés (MMC) [9][2], ces modèles permettent de spécifier explicitement des grammaires stochastiques sous un formalisme convenable.

Un MMC est un modèle de production pour des séquences. Il permet de passer d'une séquence de vecteurs (ici les mots codés) à une séquence de symboles (ici les concepts que l'on désire extraire) qui correspondra par exemple à la suite d'étiquettes la plus probable pour la séquence. Un MMC est défini par deux processus stochastiques : une chaîne de Markov définie par un ensemble d'états et les transitions entre ces états, des probabilités dites d'émission associées à chaque état. Ces dernières donneront la probabilité de générer une observation dans un état donné. Dans le cas discret, cette description est une table de probabi-

lités, et dans le cas continu, une distribution de probabilité des observations. Dans notre modélisation, les différents états des MMC codent les concepts, la structure des transitions code la «grammaire de concepts» c'est à dire les transitions entre concepts acceptables.

Pour apprendre un MMC (c'est à dire, déterminer les probabilités d'émission et de transition des modèles) nous utiliserons une base de séquences déjà étiquetés. Il existe plusieurs algorithmes d'apprentissage des MMC, nous avons utilisé l'apprentissage de Viterbi [2].

Une fois les paramètres des modèles appris, un MMC nous permet de trouver la séquence de symboles (concepts) les plus probables pour une séquence d'observations (mots) et la probabilité que la séquence d'observations ait été produite par notre MMC. Associer une séquence de concepts à une séquence de mot, permet de segmenter cette dernière par exemple en passages pertinents pour une requête et non pertinents. La valeur de la probabilité de la séquence d'observations nous permet d'autre part de juger en quoi la séquence est conforme au modèle MMC utilisé.

Plus formellement, un MMC discret est défini par un 5-uplet $\{N,M,A,B,\pi\}$, où :

– N et M sont respectivement le nombre d'états du modèle et la taille du dictionnaire,

– q_t et w_t sont l'état et la sortie du système au temps t,

– $A=\{a_{ij}\}$ est la matrice des probabilités de transition et a_{ij} est la probabilité d'aller de l'état j à l'état i : $a_{ij} = P(q_{t+1} = j \,|\, q_t = i), \quad 1 \le i, j \le N$,

– $B = \{b_j(k)\}$ est la matrice des probabilités d'émission, et $b_j(k)$ est la probabilité de produire le symbole k (noté v_k) dans l'état j : $b_j(k) = P(w_t = v_k \,|\, q_t = j), \quad 1 \le k \le M$,

– $\pi = \{\pi_i\}$ est le vecteur des probabilités d'état initiales, et π_i est al probabilité initiale de l'état i : $\pi_i = P(q_1 = i), \quad 1 \le i \le N$.

La notion de temps correspond ici à la position d'un mot dans une séquence. Un passage de texte est représenté comme une séquence de mots $\mathbf{w} = (w_1\, w_2 \ldots w_T)$, où T est la longueur du passage.

Un MMC décrit complètement un modèle de langage probabiliste. Etant donnés un MMC *M* et une séquence d'états $\mathbf{q} = (q_1\, q_2 \ldots q_T)$, la probabilité d'observer la séquence **w** est :

$$P_M(\mathbf{w}|\mathbf{q}) = \pi_{q_1} b_{q_1}(w_1) \left(\prod_{t=2,T} a_{q_{t-1} q_t} b_{q_t}(w_t) \right)$$

Les séquences d'états ne sont pas connues en générale (elle sont «cachées»). Etant donnés une séquence d'observations **w** et un MMC, ce qui nous intéressera en général est :

– la probabilité que la séquence soit produite par le modèle (la somme étant sur toutes les séquences d'états possibles) :

$$P_M(\mathbf{w}) = \sum_{\mathbf{q}} P_M(\mathbf{w}|\mathbf{q}) \quad (1)$$

– la séquence d'états la plus probable pour la séquence d'observations :

$$\mathbf{q}^*_M(\mathbf{w}) = \arg\max_{\mathbf{q}} P_M(\mathbf{w}|\mathbf{q}) \quad (2)$$

Le nombre de mots distincts présents dans le corpus est trop important pour que l'on puisse estimer les probabilités d'émission $b_j(k)$, et ceci malgré la taille importante du corpus. Nous avons choisi de coder ces mots par un codage continu sur l'axe réel. Pour cela nous avons utilisé la mesure U définie en section 2. Dans notre cas, nous utiliserons des MMC continus (des densités de probabilité remplacent les probabilités d'émission) au lieu des MMC discrets introduits précédemment dans le but de simplifier la description de ces modèles. Ce type de codage et de modélisation sont originales dans le domaine de l'EI.

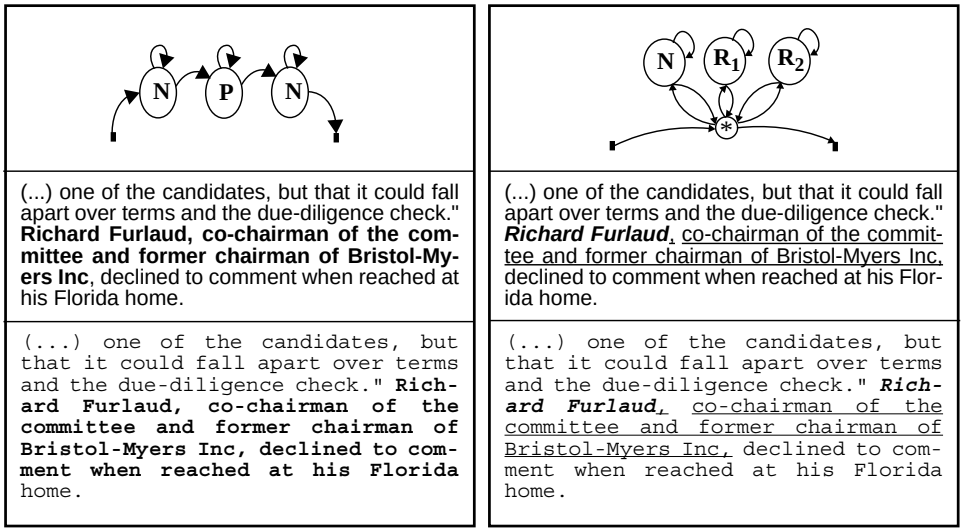

Figure 4: Deux MMCs pour l'extraction de l'information. Les modèles sont présentés en haut, l'étiquetage désiré au milieux et l'étiquetage obtenu par le modèle après apprentissage en bas. Les textes montrent une partie d'un paragraphe pertinent. Dans le modèle de gauche, le style gras indique des mots pertinents. Dans le modèle de droite, les styles souligné et gras indiquent respectivement les états POSITION (R1) et PERSONNE (R2).

A titre d'exemple, la figure 4 (gauche) représente un MMC simple, où les états sont représentés par des cercles et les transitions entre états possibles par des flèches. Ce MMC code une grammaire à deux concepts : pertinent (P) ou non pertinent (N). Si on recherche la séquence la plus probable de concepts (eq. 2) correspondant à une séquence de mots, on obtient de ce MMC une traduction de la séquence de mots codés d'un passage $\mathbf{u}=(u_1\ u_2\ ...\ u_T)$ dans la séquence de concepts $\mathbf{q}=(q_1\ q_2\ ...\ q_T)$ qui dans cet exemple ne pourra être que de la forme NPN. Le décodage (eq. 2) fournira la séquence de mots pertinents. Ce modèle peut être ainsi utilisé pour extraire le sous-passage le plus pertinent dans chaque paragraphe si la structure de ce

dernier est de la forme NPN. Si ce n'est pas le cas, la probabilité de la séquence (eq. 1) sera faible.

Nous avons utilisé ce modèle simple pour localiser des sous-passages portant sur les concepts de PERSONNE ou POSITION, sans distinguer ces deux champs. La figure 4 (gauche) donne un exemple d'un paragraphe pertinent localisé à l'étape précédente (RI) où le sous-passage pertinent est en gras. Nous donnons en dessous la segmentation de ce passage obtenue par le modèle. Sur cet exemple, la partie pertinente du passage à été détectée, mais une partie non pertinente à été également sélectionnée.

Pour pouvoir «étiqueter» l'information, il faut utiliser des modèles stochastiques plus riches, ce que nous proposons dans l'exemple suivant (figure 4, droite). Ici, nous cherchons à segmenter le texte en trois concepts : NON PERTINENT (N), PERSONNE (R1) et POSITION (R2). Contrairement à l'exemple précédent, nous avons choisi un modèle peu contraint qui traduit le fait qu'on n'a aucune connaissance a priori à propos de la grammaire des passages. La seule contrainte que nous avons imposée est de forcer les changement de concepts à s'aligner avec les début des phrases[1]. L'état «*» a donc une probabilité d'émission de 1 pour le symbole de fin/début de phrase et de 0 pour le reste. Sous le modèle nous donnons le même paragraphe que précédemment avec les segmentations désirées et obtenues. Dans l'exemple elles coïncident exactement. Ce modèle est extensible à un nombre quelconque de concepts, en conservant la même topologie.

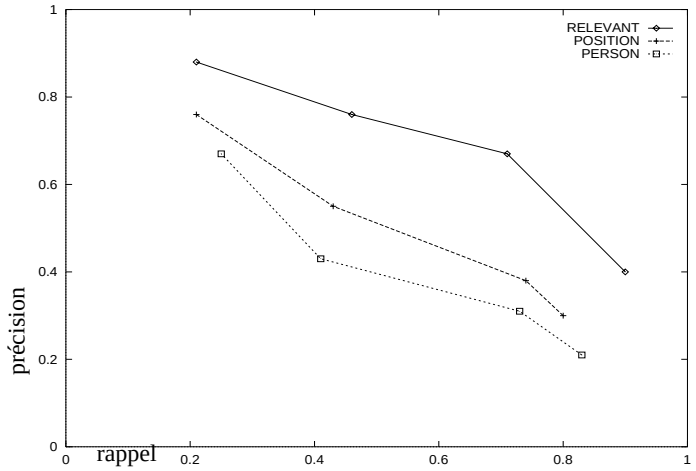

Figure 5: Performances du module stochastique d'EI.

4. Résultats

L'évaluation des systèmes en RI et EI est un problème délicat largement ouvert aujourd'hui. Nous

1. Nous définissons informellement une phrase comme une séquence d'au moins deux mots non séparés par des symboles de ponctuation.

avons choisi d'évaluer notre système d'EI par les courbes de précision-rappel déjà introduites pour la partie RI. Nous prenons cette fois le mot comme unité de comptage et non plus le document ou le paragraphe. Ainsi, si h est le nombre de mots étiquetés correctement pour un concept donné, N le vrai nombre de mots de ce concept, H le nombre total de mots étiquetés pour ce concept, alors la précision est donnée par h/H et le rappel par h/N. Pour obtenir différentes valeurs de la courbe précision-rappel ils existent plusieurs démarches [8]; ici nous avons modifié artificiellement la probabilité de transition entre états, indépendamment pour chaque concept. La figure 5 représente les courbes de précision-rappel pour chaque concept et pour les deux MMC présentés ci dessus.

Ces résultats bien que préliminaires montrent que les modèles simples que nous avons testés permettent d'extraire automatiquement des informations de surface sur des bases de textes. Ces deux modèles correspondent à deux extrêmes, le premier est très contraint, le second l'est très peu et sa structure n'incorpore aucune information *a priori* sur le texte. Une modélisation plus raisonnable devrait se situer entre les deux. Dans cet exemple, on voit également que pour ce système, les informations de type «personne» qui apparaissent dans des contextes très variés et correspondent à des séquences très courtes sont plus difficiles à retrouver que des informations correspondant à des séquences plus longues et dont les mots sont plus spécifiques du concept, comme dans le cas des informations de type «position».

5. Conclusion et Perspectives

Nous avons présenté un système d'extraction d'information basé sur des modèles d'apprentissage numérique : des classifieurs pour la localisation de l'information, et des modèles stochastiques de langage pour son analyse. L'intérêt d'une approche hiérarchique pour la localisation de l'information a été montré, du point de vue des performances et du point de vue de la qualité du codage résultant. L'utilisation de modèles Markoviens simples pour l'analyse et l'extraction de l'information a été ensuite introduite. Ces modèles ont été appliqués à une version simplifiée de la tâche *Scenario Templates* de MUC-6, qui constitue une référence dans le domaine. L'utilisation et la conception de ces modèles pour l'extraction d'information en est encore à un stade préliminaire et doit encore être développée pour devenir satisfaisante.

L'approche que nous avons proposée est originale par

– l'utilisation d'un corpus de textes et patrons comme seule base de connaissances. L'évacuation de toute connaissance linguistique explicite est très certainement une limitation mais permet de traiter de façon automatique des tâches d'extraction, avec un gain conséquent en vitesse et portabilité des applications.

– l'utilisation de la mesure U pour le codage continu des mots. Le codage des mots dans un espace de petite dimension est nécessaire pour l'utilisation de modèles d'apprentissage sur de bases de données textuelles.

– l'utilisation de modèles stochastiques pour l'analyse de textes est bien connu dans le traitement de séquences de parole ou de données biologiques, mais a été très rarement employé pour la recherche et l'extraction de l'information.

– le couplage IR - IE.

Nous nous attaquons à présent au développement de modèles stochastiques de traitement de séquences plus riches (Markoviens ou Connexionnistes) pour améliorer la reconnaissance des concepts. Nous étudions aussi des techniques de codages de mots plus évoluées. Finalement, un travail d'évaluation rigoureuse des modèles est en cours pour analyser leur comportement.

Références

[1] Andersen E. (1992) The Statistical Analysis of Categorical Data. Springer, Berlin.

[2] Charniak E. *Statistical Language Learning*, a Bradford book, The MIT Press, 1993.

[3] Grishman R. (1996) Design of the MUC-6 Evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 13-33. Morgan Kauffman.

[4] Hertz J., Krogh A. et Palmer R.G. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley Publishing Company, Redwood City, CA.

[5] Hobbs J.R. (1993) The Generic Information Extraction System. In *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Morgan Kauffman.

[6] Knaus D., Mittendorf E., Schauble P. and Sheridan P. (1995) Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System. Proc. of the 4th Text Retrieval Conference TREC-4.

[7] Koller D., Sahami M.(1996) Hierarchically classifying documents using very few words.

[8] Mittendorf E., Schauble P. (1994) Document and Passage Retrieval Based on Hidden Markov Models. In *ACM SIGIR'94*, 318-327.

[9] Rabiner L., Juang B.H. (1993) *Fundamentals of Speech Recognition,* Prentice Hall Signal Processing Series, 1993.

[10] Robertson M.A. (1997) On the Marriage of Information Retrieval and Information Extraction. *19th Annual BCS-IRSG Colloquium on IR Research* (Aberdeen 1997) 60-69 ,Furner J. Harper D. (eds.)

[11] Schutze, Hull (1995) A Comparaison of Classifiers and Dcoument Representations for the Routing Problem . In *TREC-4*, 1995.

[12] Stévenin-Barbier A. et Gallinari P. (1997) Semantic anticipation for understanding using neural networks, PACES / SPICIS (Singapore 1997).

[13] Sutton G. (1989) *Automatic Text Processing : the Transformation, Analysis and Retrieval of Information by Computer.* Addison Wesley 1989.

[14] TREC (1996) Harman D. (ed.) *Proc. of the 5th Text Retrieval Conference - TREC 5* (Washington 1996).

[15] Wiener, Pedersen (1995). A Neural Network Approach to Topic Spotting. In *Proc. of the Foruth Anual Symp. on Doc. Analysis and Information Retrieval (SDAIR'95)*, 317-321.