ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Formation Doctorale : Informatique

École Doctorale : Mathématiques et Informatique

THÈSE

pour obtenir le grade de

DOCTEUR EN INFORMATIQUE
DE L'UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

SPÉCIALITÉ : INFORMATIQUE

Présentée et soutenue publiquement le 28 septembre 2000, par

Brigitte BIGI

Contribution à la modélisation du langage pour des applications de recherche documentaire et de traitement de la parole

Composition du jury :

Mme	Michèle Jardino	CR, LIMSI, Paris	Rapporteur
MM	Jean-Paul Haton	PR, LORIA, Nancy	Rapporteur
	Marc El-Bèze	PR, LIA, Avignon	Examinateur
	Pierre Isabelle	PR, XEROX, Grenoble	Examinateur
	Renato De Mori	PR, LIA, Avignon	Directeur de thèse
	Thierry Spriet	MC, LIA, Avignon	Co-Directeur de thèse

Laboratoire d'Informatique d'Avignon

Résumé

L'application des méthodes statistiques aux domaines de la recherche documentaire et de la reconnaissance automatique de la parole (RAP) prend une importance grandissante. Ce travail de thèse présente des solutions qui utilisent des modèles de langage dynamiques, suivant la théorie de l'information. Notre contribution est l'apport de nouvelles approches en modélisation du langage. Les applications développées sont plurielles : classification thématique de textes écrits ou du discours, segmentation thématique, et expansion de requête.

En classification thématique, l'objectif est d'assigner un label thématique à un segment de texte parmi un ensemble de labels possibles. Le modèle, dans ce cas, repose sur la comparaison entre la distribution statistique des mots contenus dans la mémoire cache d'un texte à un instant donné et les distributions statiques des mots clés des thématiques. Cette évaluation évolue dans le temps avec la prise en compte de nouveaux mots dans le cache. En combinant les décisions prises par ce modèle et par un ensemble d'unigrammes thématiques classiques, on détermine le thème d'un texte avec un degré de fiabilité supérieur à 80 %. Appliqué à des textes dictés, le modèle à base de mémoire cache nous permet une reconnaissance rapide des thèmes, ce qui laisse envisager l'utilisation, dans les systèmes de RAP, d'un modèle de langage plus approprié au domaine du texte dicté. Nous montrons alors que l'utilisation d'une combinaison linéaire d'un modèle bigramme général avec des modèles thématiques apporte un gain substantiel de perplexité par rapport à l'utilisation unique d'un bigramme général et statique.

En segmentation thématique de textes écrits, on cherche à déterminer les frontières entre paragraphes de thèmes différents. Une possibilité pour repérer les changements de thèmes est d'utiliser le modèle à base de mémoire cache déjà développé pour la classification thématique. D'autres nouvelles méthodes ont également été testées. On a notamment cherché des solutions afin que la segmentation ne nécessite pas de connaissances préalables sur les thèmes, contrairement au modèle à base de mémoire cache. Pour ce faire, on donne de nouvelles représentations de l'histoire d'un mot. Les résultats obtenus sont de moindre qualité par rapport au modèle thématique, cependant l'ensemble des résultats ainsi obtenus montrent que différentes stratégies doivent être utilisées selon les valeurs de rappel et de précision que l'on souhaite.

Ce type d'approche a également été appliqué à la recherche documentaire. Le but en expansion de requête est d'ajouter de nouveaux termes pertinents à la requête d'un utilisateur afin de rendre les réponses, fournies par le système de recherche documentaire, plus précises. Notre modèle évalue la distance entre la distribution de probabilités des termes représentatifs des documents fournis par le système avec la requête initiale et la distribution de ces mêmes termes dans la collection entière. Ces évaluations permettent de donner un score à des termes candidats qui formeront la requête étendue. Ce modèle améliore uniformément les résultats par rapport à ceux obtenus avec la requête initiale, mais aussi par rapport à d'autres fonctions d'évaluation, issues de la littérature, pour l'ordonnancement des termes.

Table des matières

Introduction

L'émergence de nouvelles formes de communication a entraîné de nouveaux problèmes : le stockage, la transmission, la recherche des informations ainsi que leur utilisation. Les informations que l'on traite transitent par divers media : littérature, forum de discussion, pages web, mais aussi messages parlés, émissions de radio... Cependant leur forme reste classique, il s'agit surtout de données textuelles, de messages audio, d'images ou de vidéos, qu'il faut pouvoir classer pour accéder plus efficacement à cette masse toujours grandissante d'informations. Un traitement manuel est, aujourd'hui, devenu illusoire face à la quantité de données. Les méthodes automatiques de classification et de segmentation s'imposent alors d'elles-mêmes.

Les corpora consistent ainsi en un flot de données continues incluses dans des segments de nature différente. La segmentation et la classification consistent à séparer des textes en entités de nature proche, et à regrouper automatiquement ces entités en classes en assignant une ou plusieurs de ces classes à chaque portion de texte. La classification a pour but de ranger les données dans des catégories selon un certain ordre et suivant une certaine méthodologie. Ainsi, plusieurs types de classifications peuvent être réalisées (par type de message, par genre littéraire, par thématique abordée...) et selon diverses méthodes (statistique, par similarité, ...). Les travaux présentés ici s'intéressent à la classification thématique de textes par des méthodes statistiques. L'objectif est d'assigner un label thématique à une entité textuelle, parmi un ensemble de labels possibles.

La segmentation, et notamment selon un critère thématique, est un processus utilisé dans de nombreuses applications. En recherche documentaire, segmenter les documents fournis en réponse à la requête d'un utilisateur permet de ne lui proposer que les segments pertinents. Dans ce cas, on considère, par exemple, que l'utilisateur ne désire consulter que la partie d'un document qui contient des occurrences des mots exprimés dans la requête. Ces "micro-documents" peuvent aussi être créés selon des critères de cohésion lexicale présents dans le texte (répétition des termes et relations sémantiques), ou par d'autres méthodes comme les arbres de décision ou les méthodes statistiques.

Cette thèse s'inscrit dans le cadre de l'analyse statistique du langage pour les domaines de la recherche documentaire et de la reconnaissance automatique de la parole. Les méthodes statistiques sont employées en recherche documentaire par exemple pour l'indexation de documents écrits et audio, notamment grâce à leurs bonnes performances. Les méthodes fondées sur des modèles statistiques sont également largement utilisées dans les systèmes de reconnaissance automatique de la parole.

Le premier chapitre introduit la reconnaissance automatique de la parole (RAP) et les modèles de langage. La RAP a pour objet la transformation automatique d'un signal acoustique en une séquence de mots, qui, idéalement, correspond à la phrase prononcée par un locuteur. Les systèmes couramment utilisés sont constitués de deux composantes : acoustique et linguistique. La première composante procède à un décodage acoustique qui a pour objet de transformer le signal, reçu par l'intermédiaire d'un microphone, en un ensemble d'hypothèses de mots auxquelles on attribue une localisation temporelle dans la phrase. La seconde composante, qui incorpore un modèle de langage, évalue pour chaque mot du vocabulaire sa probabilité d'apparition, étant donnée une liste de mots candidats élaborée précédemment. La composante linguistique ordonne les séquences de mots en fonction de la probabilité de leur apparition dans le langage. L'objectif principal des modèles stochastiques du langage est de donner une information permettant de guider la reconnaissance par une définition des séquences de mots considérées comme correctes ou probables.

Le deuxième chapitre aborde le domaine de la recherche documentaire. Les systèmes de recherche documentaire prennent en entrée la question formulée par un utilisateur et y répondent en proposant une série de documents appartenant à un corpus, qui correspond au mieux à la requête formulée. Ce processus diffère de l'interrogation classique de bases de données car les questions sont formulées par les utilisateurs qui ne connaissent pas le langage formel utilisé par la base. De plus, l'utilisateur n'en connaît pas les structures. Effectuer une correspondance entre la question d'un utilisateur et l'ensemble des documents connus est donc une opération délicate dont le résultat peut comporter de nombreuses erreurs, notamment lorsque la requête de l'utilisateur est courte. Afin de rendre la recherche plus efficace, il est possible d'utiliser une classification hiérarchique des documents incluant une indexation thématique. Une autre solution consiste à segmenter les documents et comparer les requêtes avec des segments plus homogènes que les documents entiers.

Le troisième chapitre présente notre approche dans le domaine de la classification thématique de textes écrits et du discours. Le problème est l'assignation d'un label thématique à un segment de texte parmi un ensemble de labels possibles. Le modèle que nous présentons combine une mémoire cache avec un modèle de backing-off pour les n-grammes. Il utilise une sélection de mots clés pour donner une représentation de chacun des thèmes. A chaque instant, une mesure de Kullback-Liebler compare, pour tous les thèmes, la distance entre la distribution statistique des mots dans la mémoire cache et la distribution des mots clés du thème. L'ensemble des distances ainsi obtenues permettra de déterminer dynamiquement le thème du texte traité. Ce modèle est également associé à un ensemble d'unigrammes thématiques.

Le quatrième chapitre décrit le problème de l'adaptation des modèles de langage qui permet de capter les changements de nature du texte : changement de domaine, changement de sujet, de thématique... Nous nous intéressons plus particulièrement à l'évaluation des paramètres des modèles de langage statistiques et à l'adaptation dynamique. Nous proposons d'évaluer les performances des modèles de langages thématiques, en vue de leur intégration dans un système de RAP. L'objectif est de sélectionner dynamiquement le modèle thématique le plus pertinent. Cette approche conduit à une réduction substantielle de la perplexité par rapport à celle obtenue avec un modèle classique.

Dans le cinquième chapitre, nous étudions le problème de la segmentation thématique. Nous proposons de le décomposer en deux étapes : la première vise à déterminer un ensemble de ruptures thématiques candidates et par la suite une phase de sélection détermine leur pertinence. Plusieurs méthodes sont proposées, dont certaines nécessitent l'utilisation de connaissances *a priori* sur les thèmes, tandis que d'autres en sont indépendentes. Nous repérons les changements de thèmes avec le modèle à base de mémoire cache défini pour la classification thématique et avec une méthode basée sur des bigrammes à distance combinés avec une mémoire cache. Concernant la sélection des candidats, nous comparons le modèle à base de mémoire cache avec une méthode où l'on recherche une séquence optimale de candidats en faisant varier la taille de l'historique.

Dans le dernier chapitre, nous proposons une solution pour limiter le nombre de documents fournis par les systèmes de recherche documentaire. L'expansion de requête permet de restreindre les erreurs de correspondance entre la question d'un utilisateur et les documents proposés par le système et donc de réduire le nombre de documents ciblés. Le principe est d'augmenter la requête avec de nouveaux mots ou expressions dont la signification est proche. Nous proposons de réaliser automatiquement l'expansion de requête par retour de pertinence. Le principe est d'utiliser les documents fournis comme pertinents à partir de la requête initiale et d'en extraire automatiquement un ensemble de nouveaux termes. On propose d'évaluer la pertinence des termes candidats en calculant la distance entre la distribution des meilleurs termes des documents fournis par le système de recherche à partir de la requête initiale et la distribution de ces termes dans la collection entière des documents. Cette mesure est évaluée avec la divergence de Kullback-Leibler, issue de la théorie de l'information.

Chapitre I

Reconnaissance Automatique de la Parole

1 Quelques notions de théorie de l'information

On peut voir la communication Homme-Machine comme un échange d'informations à travers un moyen physique. Ces informations sont codées avec une méthode appropriée pour la transmission ([De Mori 1998]). Le processus qui produit une représentation de ce qui doit être communiqué s'appelle le codage. Le contenu de ce qui doit être communiqué est structuré en utilisant des mots représentés par des séquences de symboles d'un alphabet et selon un lexique donné. La concaténation des mots selon les règles d'une grammaire donne des expressions. Ces types variés de contraintes sont des sources de connaissances ([Junqua et Haton 1996]) avec lesquelles on construit une version symbolique du message à échanger. Celle-ci subit plusieurs transformations qui rendent le message transmissible sur le canal physique. Ce type de communication s'appuie sur la théorie de l'information dont cette section présente les principes de base.

La théorie de l'information a vu le jour avec le développement croissant des télécommunications. L'un des problèmes est de transmettre efficacement les informations, en fonction des différentes contraintes, notamment celles dues aux propriétés de la ligne. La théorie de l'information fournit une mesure quantitative de la notion d'information apportée par un message (ou une observation). Cette notion fut introduite par Claude Shannon en 1948 afin de répondre à deux questions fondamentales dans le domaine des techniques de communication : quelle est la limite en matière de compression de données - l'entropie H du message, et quel est le débit maximal de transmission d'informations au moyen de canaux bruités - la capacité C du canal ([Shannon 1948]).

La théorie de l'information, telle qu'elle fut élaborée par Shannon avait pour objectif principal d'évaluer les performances limites des systèmes de télécommunications *en présence de perturbations aléatoires*. La figure I.1 représente le schéma de communication désigné sous le nom de *paradigme de Shannon*. Une *source* engendre un *message* à l'intention d'un *destinataire*. La source et le destinataire sont deux entités séparées qui sont reliées par un *canal* qui est le support de la communication d'une part, mais qui d'autre part est le siège de *perturbations*. Ces dernières, aleátoires, ont pour effet de créer une différence entre le message émis et celui qui est reçu. Par ailleurs, le message émis par la source est, dans une certaine mesure, imprévisible, car s'il était connu à l'avance la communication serait inutile.

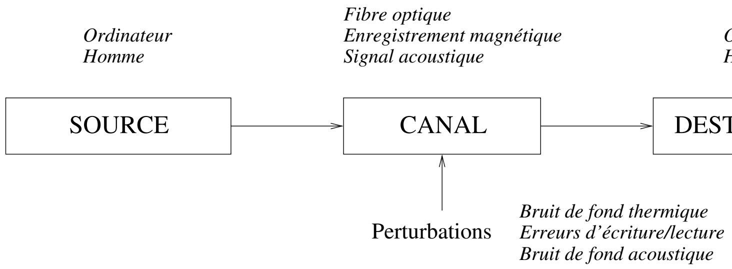


FIG. I.1 – Schéma fondamental d'une communication

Un canal d'information se décrit en un alphabet donné d'entrée $A = \{a_i\}$, $i = 1, 2, \cdots, r$, un alphabet de sortie $B = \{b_j\}$, $j = 1, 2, \cdots, s$ et un ensemble de probabilités conditionnelles $P(b_j \mid a_i)$ pour tous les i et j possibles. $P(b_j \mid a_i)$ est la probabilité que le symbole de sortie b_j soit reçu quand a_i a été émis.

2 Principes de reconnaissance

Les systèmes de reconnaissance automatique de la parole accomplissent un processus de décodage (voir [Jelinek 1976], [Jelinek et Mercer 1980], [Bahl et al. 1983], [Dugast et al. 1995], [Jelinek 1997]) en utilisant les sources de connaissance pour transformer le message représenté par un signal de parole avec différents niveaux de représentations symboliques. Le décodage peut produire des séquences de mots ou des hypothèses conceptuelles.

Contrairement à la communication Homme-Homme, la communication Homme-Machine doit produire des instances de données structurées de façon déterministe, ce qui signifie que le système informatique doit engendrer la même représentation chaque fois qu'un même signal est traité. Les sources de connaissance utilisées durant le décodage par les machines sont seulement des *modèles* de celles utilisées par l'homme pour interpréter ses messages. L'une de ces connaissances est le *modèle de langage* qui représente le regroupement de contraintes sur les séquences de mots acceptables dans un langage donné ([Federico et al. 1995], [Kneser et Steinbiss 1993], [Kuhn et De Mori 1990]).

Une des difficultés importantes de l'analyse du langage naturel est qu'il est presque impossible de concevoir une grammaire **G** capable de générer à la fois toutes les phrases du langage, mais aussi seulement celles-ci. Ceci est dû à de nombreux facteurs dont le plus important est probablement le fait que les langages naturels évoluent et qu'il est difficile de les caractériser avec des modèles formels. Les grammaires qui ont un grand nombre de règles détaillées peuvent modéliser avec précision certains aspects du langage mais sont trop limitées pour certains autres. Ces grammaires sont dites à faible couverture.

D'autres grammaires peuvent avoir une couverture complète mais, en restant très générales, elles peuvent générer des phrases qui n'appartiennent pas au langage. Un exemple de ce type de grammaire sur-productrice est celle qui peut générer toutes les paires de mots du vocabulaire du langage (*word pair grammars*). La surproduction peut être atténuée par l'association de probabilités aux règles de la grammaire dans laquelle les phrases indésirables seront générées avec des probabilités inférieures à celles des phrases correctes du langage. Certaines de ces grammaires sont particulièrement utilisées pour la reconnaissance automatique de la parole car elles peuvent être représentées par des automates stochastiques à nombre d'états fini où les probabilités sont associées aux arcs.

Les arcs dans ces automates stochastiques à état-fini peuvent alors être remplacés par d'autres automates, un pour chaque mot, et dans ce cas, ils représentent les prononciations de chacun des mots. Les arcs de ces automates de mots sont étiquetés avec des phonèmes et les prononciations sont obtenues avec un *modèle lexical*. Chaque modèle de phonème peut être remplacé par un *modèle acoustique* qui est un automate avec des arcs associés à des distributions de paramètres acoustiques ou des caractéristiques qui peuvent être observées.

A partir de cette hiérarchie de composants, un réseau intégré peut être obtenu et effectivement utilisé pour générer des hypothèses sur les mots ou les interprétations à propos d'un signal de parole donné.

Le processus de décodage doit traiter les *ambiguïtés*. Celles-ci peuvent avoir plusieurs origines comme les distorsions introduites par le canal de transmission, ou les différentes façons de prononcer et articuler un mot, ou encore celles dues aux imprécisions intrinsèques du message énoncé. Les ambiguïtés peuvent être réduites en exploitant les *redondances* du message. En pratique, la connaissance est utilisée pour transformer le signal d'entrée en plusieurs séquences de vecteurs de paramètres et pour obtenir à partir de ceux-ci différents niveaux de représentation symbolique. Le premier de ces niveaux peut être le mot, la syllabe, le phonème ou simplement un descripteur acoustique.

L'interprétation est habituellement obtenue par un processus de *recherche* qui considère un réseau intégré comme étant le générateur d'une description observable $X = x_1x_2...x_n...x_N$ d'un signal à analyser. Le processus de recherche tente de trouver la meilleure séquence des états du réseau intégré en recherchant un chemin qui génère X. L'obtention des séquences candidates s'appuie sur des méthodes d'évaluation dont les résultats sont des scores utilisés par des stratégies de recherche pour déterminer progressivement les chemins partiels du réseau intégré.

Les systèmes modernes sont basés sur des scores probabilistes attribués aux hypothèses de candidats. Un modèle probabiliste simple pour évaluer les hypothèses a une connaissance du décodeur qui considère une séquence d'observations acoustiques comme la sortie d'un canal d'information (Figure I.2) qui reçoit en entrée une séquence de symboles représentant l'intention du locuteur. Si ces symboles sont des mots $W = w_1w_2 \cdots w_k \cdots w_K$ on a X qui est la version codée de W. Le but de la reconnaissance est de reconstruire W en se basant sur l'observation de X. Ceci est réalisable en utilisant la connaissance du processus de codage. Si le même X peut être produit par différents W, ou si la connaissance est incomplète ou imparfaite, alors la reconstruction peut ne pas réussir.

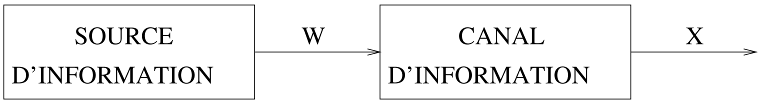


FIG. I.2 – Un modèle simple de décodeur

Dans le cas de la dictée vocale, l'ambiguïté et l'imprécision impliquent la nécessité de considérer la reconnaissance comme un processus de recherche qui génère des hypothèses de mots par la sélection de candidats pour lesquels $P(W \mid X)$ est maximale. Si le modèle de la source fournit $P(W)$ et le modèle du canal $P(X \mid W)$, alors $P(X,W) = P(X \mid W)P(W)$ peut être calculée. Il faut noter que, comme $P(X)$ est la même pour tous les candidats W, la séquence W' pour laquelle $P(X,W)$ est maximale est la séquence pour laquelle $P(W \mid X)$ est maximale.

$P(X \mid W)$ est la probabilité d'observer X quand W est prononcé. En pratique, cette probabilité ne peut pas être évaluée directement à partir des données. Elle doit l'être avec un modèle acoustique. $P(W)$ est la probabilité d'une séquence de mots et elle s'évalue avec un modèle de langage. La contribution d'un modèle de langage (ML) aux systèmes de RAP est basée sur la plausibilité linguistique de l'occurrence d'un mot, immédiatement après les mots précédemment reconnus.

3 Modèles de langage

Malgré le recours à des modèles acoustiques complexes, la phrase reconnue par un système automatique peut être partiellement erronée. Par exemple, la phrase *"L'équipe de hockey sur glace"* pourrait être reconnue *"L'équipe de OK sur glace"*. L'introduction de contraintes sur les séquences de mots a pour effet de restreindre la reconnaissance à des phrases syntaxiquement correctes ou à des séquences probables, diminuant ainsi le taux d'erreurs. Différents types de modèles de langage ont été proposés dans le but de restreindre les séquences de mots envisagées au cours d'une reconnaissance. Les modèles les plus classiques sont présentés dans cette section.

L'essentiel de l'information contenue dans un modèle de langage relève de la notion d'énoncés plus ou moins conformes de la langue naturelle. Les modèles à base de connaissance se différencient en plusieurs points des modèles statistiques. Un modèle de langage à base de connaissances syntaxiques peut ainsi renseigner sur le fait qu'une phrase est grammaticalement bien construite. Un autre modèle, pour peu qu'il possède des connaissances sémantiques sur le contexte dans lequel est employée la phrase, pourra renseigner en plus sur le fait que la phrase a un sens. Mais ces modèles fournissent une distinction correct/incorrect, sans nuances. Les modèles statistiques, pour leur part, présentent l'avantage de fournir une réponse quantifiée, en termes de probabilités, à la conformité des phrases par rapport à un langage.

La principale utilisation des modèles de langage est la reconnaissance automatique de la parole continue (récemment présentée dans la thèse de G. Damnati [Damnati 2000]). Dans ce contexte, ils permettent de guider le processus de reconnaissance en fournissant des informations sur la plausibilité des hypothèses retenues ou en cours d'exploration. Les modèles de langage prennent également une part importante dans le traitement du langage écrit, notamment en recherche documentaire (chapitre II). Nous présenterons des méthodes de classification (chapitre III), des méthodes de segmentation (chapitre V) et d'expansion de requête (chapitre VI).

3.1 Les modèles n-grammes

Un modèle de langage a pour but de capter, d'adapter et d'exploiter les caractéristiques du langage naturel. Il a un impact important sur les performances des systèmes pour lesquels il est dédié. Un modèle statistique du langage permet d'attribuer une probabilité P à une séquence de mots W. Il est caractérisé par des distributions de probabilités dont les paramètres spécifient le modèle. L'objectif du modèle de langage est d'évaluer au mieux la probabilité $P(W_1^I)$ d'une séquence de mots $W_1^I = w_1 \ldots, w_i, \ldots, w_I$. La probabilité qu'un mot w_i apparaisse dans une séquence de mots est conditionnée par son historique. La probabilité d'une suite de mots peut alors s'exprimer comme le produit des probabilités de chaque mot étant donné les mots qui le précèdent, telle que :

$$P(W_1^I) = \prod_{i=1}^{I} P(w_i \mid h_i) \tag{I.1}$$

où $h_i = w_1, \ldots, w_{i-1}$ peut être considéré comme l'*histoire* du mot w_i. h_i, pour chaque w_i est apprise à partir d'un corpus d'apprentissage. Mais, au fur et à mesure que la séquence de mots h_i s'enrichie, les probabilités $P(w_i \mid h_i)$ sont de plus en plus difficiles à estimer, car il devient de moins en moins probable que la suite de tous les mots de h_i ait fait partie des données de l'apprentissage.

Afin de réduire la complexité du modèle de langage, et par conséquent de son apprentissage, les systèmes utilisent un historique réduit. Ces modèles sont dits n-grammes. La langue française dont la majorité des contraintes sont des contraintes de proximité se prête bien à ce genre de modèles. En pratique, en fonction de la taille du corpus et de l'application visée, nous utilisons des tailles d'historique de un, deux ou trois mots (unigrammes, bigrammes ou trigrammes) tels que :

$$\begin{array}{ll} \textit{Unigramme :} & P(W) = \prod_{i=1}^{I} P(w_i) \\ \textit{Bigramme :} & P(W) = \prod_{i=1}^{I} P(w_i \mid w_{i-1}) \\ \textit{Trigramme :} & P(W) = \prod_{i=1}^{I} P(w_i \mid w_{i-2}, w_{i-1}) \end{array}$$

Les modèles proviennent d'une phase d'apprentissage qui consiste à estimer les comptes c (le nombre d'occurrences) des unigrammes, bigrammes ou trigrammes. Une estimation des fréquences relatives d'apparition d'un mot selon son historique est ainsi obtenue. Par exemple, pour un modèle trigramme :

$$f(w_i \mid w_{i-2}, w_{i-1}) = \frac{c(w_i, w_{i-1}, w_{i-2})}{c(w_{i-1}, w_{i-2})}$$

Ces fréquences relatives correspondent à un maximum de vraisemblance des probabilités. Cette estimation est évaluée selon un vocabulaire de référence V. Tous les mots n'appartenant pas au dictionnaire sont regroupés sous le compte d'un jeton particulier que l'on peut noter $MOTINC$.

L'approximation n-gramme réduit considérablement les données statistiques à collecter pour évaluer $P(W_1^I)$ (voir notamment dans [Brown et al. 1992], [Della Pietra et Della Pietra 1994], [Lau et al. 1993], [Merialdo 1994], [Rosenfeld 1994]). Evidemment, une telle approximation cause une réduction de la précision. Cependant, même un modèle 3-gramme (trigramme) requiert une très grande quantité de données (corpus de textes) pour estimer efficacement le modèle. Par exemple, un trigramme de vocabulaire de taille 1 000 a besoin de l'estimation de 10^9 probabilités. Un autre aspect important, inhérent aux corpora, qui montre la difficulté de l'estimation des n-grammes est le manque de couverture et le manque de qualité des données. Dans les expériences, certaines séquences correctes de mots sont rares et n'apparaissent ainsi que peu de fois, sauf, si le corpus est très grand. La validité d'une estimation des paramètres du modèle est ainsi fortement conditionnée par la quantité d'observations disponibles.

Un autre aspect important des ML est le choix de la taille du vocabulaire. En fait, une augmentation de la taille du vocabulaire provoque un accroissement considérable du nombre de n-grammes et par conséquent leur apprentissage nécessite une masse de données et de ressources de calculs nettement supérieure. Ceci est évité en limitant le vocabulaire aux mots les plus fréquents du corpus. Néanmoins, ce choix provoque un certain taux de mots hors vocabulaires (MHV).

3.2 Les modèles n-classes

Une approche permettant de réduire la complexité des modèles de langage est de regrouper les mots en classes selon des comportements syntaxiques ou sémantiques proches, ou en classes obtenues par classification automatique ([Steinbiss et Kneser 1997]). L'espace de paramètres requis par les modèles n-grammes peut ainsi être réduit, et par conséquent la fiabilité des évaluations peut être améliorée. En effet, puisque le nombre de classes est inférieur au nombre de mots, il est plus réaliste de rencontrer beaucoup plus souvent les séquences de classes que les séquences de mots correspondantes. On notera C_i le nom de la classe du mot w_i.

La probabilité d'apparition d'un mot à la suite d'un historique dépend de la classe à laquelle appartient ce mot et de la probabilité d'apparition de cette classe à la suite de l'historique de ce mot. Si les classes sont des classes d'équivalence (un mot n'appartient qu'à une seule classe), les probabilités triclasses s'écrivent :

$$P(W) = \prod_{w_i, C_i} P(w_i|C_i) P(C_i|C_{i-2}, C_{i-1})$$

avec la probabilité suivante d'appartenence d'un mot à une classe :

$$P(w_i|C_i) = f(w_i|C_i) = \frac{c(w_i, C_i)}{c(C_i)}$$

où $c(.)$ est le nombre d'occurrences de l'argument dans le corpus d'apprentissage. La probabilité de succession de classes est définie telle que :

$$P(C_i|C_{i-2}, C_{i-1}) = \lambda . f(C_i|C_{i-1}, C_{i-2}) + (1-\lambda) . f(C_i|C_{i-1}) + \varepsilon$$

et λ est estimé automatiquement pour chaque C_{i-1} lors de l'apprentissage.

Pour réaliser de telles estimations, il est nécessaire que le corpus d'apprentissage soit étiqueté, c'est à dire que l'on affecte une classe à chacun des mots avant d'apprendre les modèles. Il existe plusieurs méthodes (manuelles ou automatiques) pour l'étiquetage d'un texte ([Church 1988], [Spriet et El-Bèze 1998]).

Plusieurs variantes des modèles n-classes sont présents dans la littérature. L'un des plus connus est le modèle POS *Part Of Speech*, qui se fonde sur une partition de parole. Dans ces modèles, un mot peut appartenir à plusieurs classes à des instants différents. Par définition, chaque occurrence d'un mot doit avoir une seule classe à un instant donné. En pratique, il est difficile d'extraire la vraie classe d'un mot dans un contexte, parmi l'ensemble des classes associées à ce mot. Ainsi, pour estimer la probabilité d'un mot w_i dans un contexte, le modèle POS calcule la somme des probabilités n-classes sur toutes les classes associées au mot à prédire.

Néanmoins ces modèles (n-classes et POS) ont le même inconvénient que celui des modèles n-grammes, à savoir la taille limitée de l'historique pris en compte. Plusieurs types de modèles qui apportent une solution à ce problème seront évoqués au chapitre IV. Par ailleurs, une étude de ce problème a été conduite récemment, dans la thèse d'I. Zitouni ([Zitouni 2000]).

3.3 Mesures d'évaluation des modèles de langage

En général, les modèles de langage pour la reconnaissance de la parole sont évalués en fonction de leur impact sur la précision de la reconnaissance. Néanmoins, ils peuvent être évalués séparément si l'on considère, par exemple, leur capacité de prédiction des mots d'un texte ([Jelinek 1990]). La mesure la plus utilisée est la *perplexité*. Mais, même si la perplexité est généralement un bon indicateur de la qualité du modèle, sa corrélation avec le taux de reconnaissance n'est pas certaine. En réalité, la qualité de la reconnaissance est influencée par la similarité acoustique des mots, ce qui n'est pas pris en compte dans les évaluations de perplexités. Une autre mesure de la qualité des modèles de langage s'appuie sur le Jeu de Shannon ([Shannon 1951]). Cette section aborde tout d'abord les notions de théorie de l'information sous-jacentes au calcul de la perplexité en tant que mesure d'évaluation des modèles, puis elle présente l'évaluation des modèles par le Jeu de Shannon.

Quantifier l'information :

La mesure quantitative de l'information est exclusivement basée sur la notion d'imprévisibilité d'un événement. L'observation d'un événement apporte d'autant plus d'information que celui-ci est improbable. Inversement, si cet événement est *a priori* très probable, son information sera faible.

L'*information propre* apportée par l'observation d'un événement a est définie par :

$$h(a) = -\log P(a) = \log \frac{1}{P(a)}$$

L'*information conditionnelle* $h(b \mid a)$ mesure l'information fournie par l'observation de b qui n'était pas déjà fournie par a :

$$h(b \mid a) = -\log P(b \mid a) = \log \frac{1}{P(b \mid a)}$$

Entropie

L'information quantitative associée à un événement est essentiellement liée au caractère plus ou moins incertain de celui-ci. La quantité d'information augmente avec l'incertitude et on utilise souvent de façon interchangeable les deux termes *incertitude* et *quantité d'information*. La quantité d'information peut être vue comme une mesure de la réduction de l'incertitude, suite à l'observation d'un événement. L'entropie mesure l'incertitude moyenne d'une variable aléatoire discrète. L'incertitude d'un ensemble de symboles *indépendants* est la somme des incertitudes des composantes. Le nom d'entropie se justifie par l'identité de sa définition mathématique avec l'entropie thermodynamique.

La mesure de Shannon établit la quantité d'information moyenne associée à chaque symbole de la source sans mémoire. Elle est définie comme l'espérance mathématique de l'information propre fournie par l'observation de chacun des symboles possibles de A. Elle est notée :

$$H(A) = -\sum_{A} P(a) \log P(a)$$

où $P(a)$ sont les probabilités des r différentes valeurs de A. Cette valeur d'incertitude est appelée entropie de la source. Il s'agit de l'information qu'on obtiendrait en moyenne en observant en parallèle les symboles émis par un très grand nombre de sources sans mémoire identiques.

Information mutuelle

L'information mutuelle est égale au nombre moyen de bits nécessaire pour spécifier un symbole d'entrée avant de recevoir un symbole de sortie moins le nombre moyen de bits nécessaires pour spécifier un symbole d'entrée après réception d'un symbole de sortie. L'information mutuelle, symbolisée dans la figure I.3, s'écrit donc :

$$i(A;B) = H(A) - H(A \mid B) \tag{I.2}$$

où :

$$H(A \mid B) = -\sum_{A,B} P(a,b) \log P(a \mid b)$$

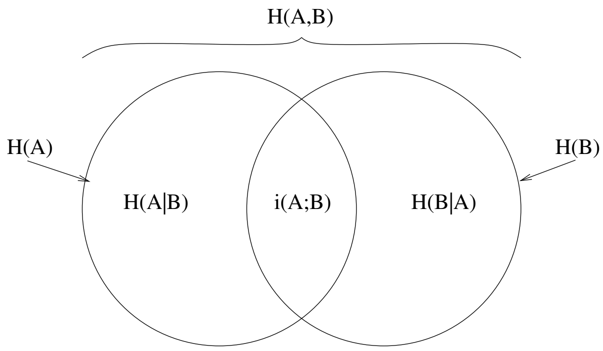


FIG. I.3 – Relations entre quelques quantités concernant les canaux

L'information mutuelle est considérée comme une mesure quantitative fournie par A sur B. Elle est donc définie comme étant la réduction apportée à l'entropie de B sachant que A est connu. Une autre façon d'écrire l'information mutuelle est :

$$i(A;B) = \sum_{A,B} P(a,b) \log \frac{P(a \mid b)}{P(a)P(b)} \tag{I.3}$$

Perplexité

Dans le cadre de la reconnaissance de la parole, si L est la taille du vocabulaire V et si la génération des mots est indépendante de la génération de leur histoire, la mesure de l'information (l'entropie) est alors définie de la façon suivante :

$$H(S) = -\left\{\sum_{i=1}^{L} P(w_i)\log(P(w_i))\right\}$$

Une source d'entropie H délivre le même contenu d'information qu'une source ayant des mots équiprobables dans l'alphabet V' de taille L' tel que $L' = 2^H$. Si la source génère des séquences de symboles de longueurs n, alors :

$$H(S) = -\lim_{n\to\infty}\frac{1}{n}\left\{\sum_{W_1^n} P(W_1^n)\log(P(W_1^n))\right\}$$

où la somme est étendue à toutes les séquences possibles de mots W_1^n.

Si la source est ergodique et que ses propriétés statistiques ne varient pas dans le temps, alors, toutes les séquences de même longueur ont la même probabilité et l'entropie devient :

$$H(S) = -\lim_{n\to\infty}\frac{1}{n}\log\{P(W_1^n)\}$$

Quand l'entropie est estimée en utilisant un corpus de n mots, on obtient alors :

$$H(S) = -\frac{1}{n}\log\{P(W_1^n)\}$$

Lorsque les probabilités sont évaluées avec un modèle, il est fréquent que l'on utilise plutôt une quantité **logprob**. Dans le cas d'un modèle bigramme de mots, on la définit de la façon suivante :

$$LP(W_1^n) = -\frac{1}{n}\log\{P'(W_1^n)\} = -\frac{1}{n}\log\left\{P(w_1)\prod_{i=2}^{n} P(w_i) \mid w_{i-1})\right\} \tag{I.4}$$

La perplexité d'un modèle de langage dérivée d'un corpus est définie comme suit :

$$PP = 2^{LP(W_1^n)} = P(W_1^n)^{-\frac{1}{n}} \tag{I.5}$$

Pour l'évaluation d'un modèle de langage, on estime les probabilités du modèle avec un ensemble d'apprentissage et on évalue la perplexité avec ce modèle sur un corpus de texte entièrement différent du corpus d'apprentissage. Ceci montre que la perplexité est très dépendante des données d'apprentissage.

Jeu de Shannon

La méthode nommée "Jeu de Shannon" permet d'estimer l'entropie d'un langage. C. Shannon l'énonce ([Shannon 1951]) de la façon suivante : une personne doit deviner la première lettre d'un texte en proposant successivement plusieurs candidats parmi les 27 possibilités[1], tant qu'elle n'a pas la bonne réponse. Une fois que la première lettre est trouvée, elle doit deviner la deuxième, puis la troisième, etc... Shannon relie les statistiques du nombre d'essais pour parvenir à la bonne réponse à la distribution des fréquences des rangs de la réponse correcte.

[1] 26 lettres de l'alphabet en anglais et l'espace

Les modèles de langage sont généralement évalués par le calcul de la perplexité sur un corpus de test. L'obtention des valeurs prises par ce critère découle de la fonction de vraisemblance. M. El-Bèze, F. Bimbot et M. Jardino ont proposé une approche alternative pour le calcul de la perplexité ([El-Bèze et al. 1997], [Bimbot et al. 1997], [Jardino 1998]). La méthode est une variante du jeu de Shannon en ce sens qu'il faut miser sur le premier mot manquant d'une phrase tronquée. Selon cette approche, un ensemble de phrases tronquées est sélectionné. Pour chacune de ces phrases, il faut répartir un capital d'une valeur de 1 entre les différents mots du vocabulaire, c'est à dire qu'il faut parier sur le mot qui suit. La perplexité est ensuite évaluée comme l'inverse de la moyenne géométrique des mises placées sur le mot correct.

Chapitre II

Recherche Documentaire

1 Principes de recherche

Le processus général d'un système de recherche documentaire est représenté dans la figure II.1. L'utilisateur pose une question en (pseudo-)langage naturel au système et celui-ci lui fournit un ensemble de documents classés selon leur pertinence par rapport à la requête formulée. Effectuer la correspondance entre la question d'un utilisateur et l'ensemble des documents est une opération délicate souvent imparfaite. Cette correspondance est réalisée par un système de recherche qui renvoie une liste triée de documents candidats. Une introduction au problème et aux algorithmes est présentée dans [Salton et McGill 1983].

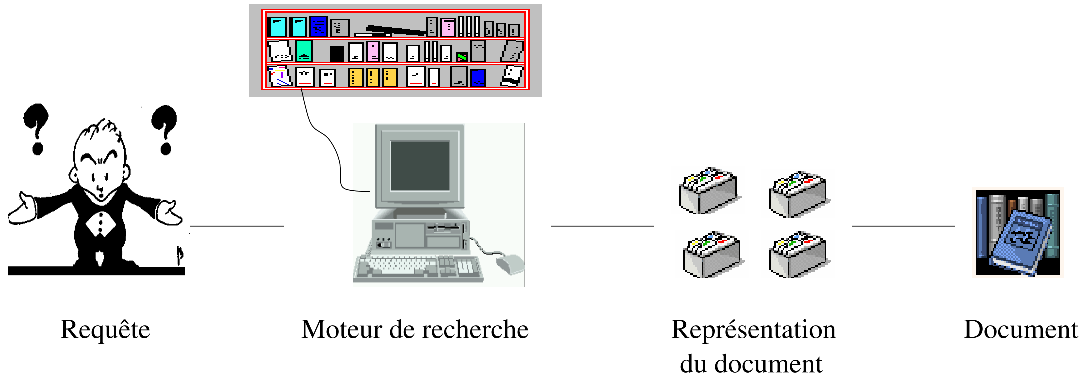


FIG. II.1 – Processus général de recherche dans un système de recherche documentaire

Les systèmes classiques de recherche documentaire représentent la question Q de l'utilisateur sous la forme d'un ensemble de mots clés qu'ils comparent avec ceux qui représentent les documents. Le résultat de cette correspondance est une liste de documents qui sont proposés à l'utilisateur. L'ensemble de réponses peut être plus ou moins précis selon le système que l'on utilise. La figure II.2 donne une représentation plus détaillée du processus de recherche. Pour une étude complète des algorithmes en jeu se référer à [Van Rijsbergen 1979], [Salton et McGill 1983], [Kowalski 1997], [Wong et Yao 1995].

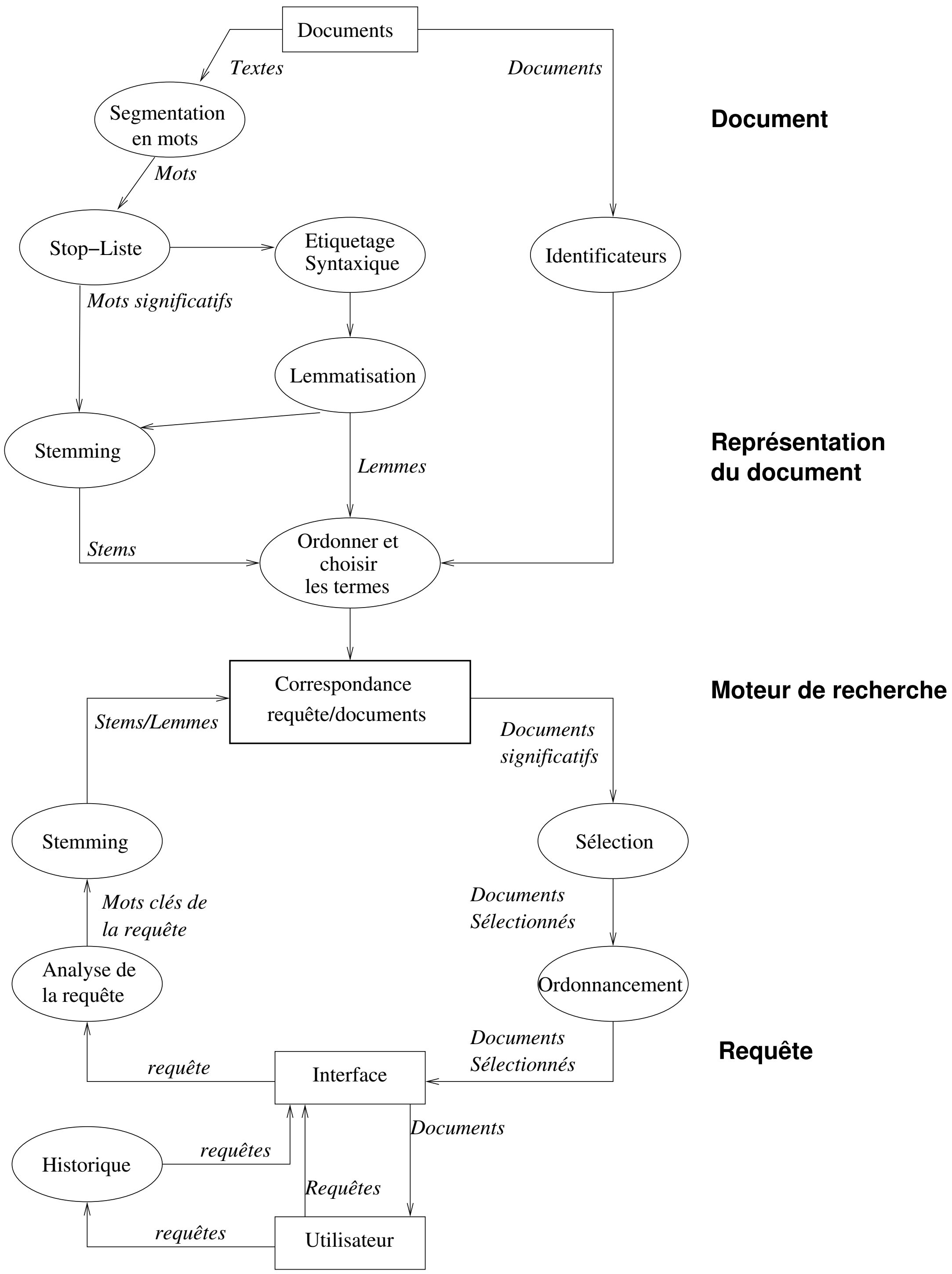


FIG. II.2 – Processus de recherche documentaire de textes

Le processus à réaliser pour traiter un document consiste à lui assigner un identificateur de telle sorte qu'il puisse être enregistré et recherché dans une base de donnée. Dans le cas de textes, obtenir une information qui se substituera au document implique sa segmentation en mots, en ignorant les mots non représentatifs par l'utilisation d'une stop-liste[2], et en obtenant les stems pour chaque mot et enfin en estimant les poids à affecter aux termes dans le document basé sur les statistiques des stems notamment (une comparaison, avec ou sans stem est présentée dans [De Loupy et al. 1999]). Le "stem" est une partie d'un mot dont le but est de définir ce mot par une de ses parties qui le caractérise, lui, et tous les termes qui lui sont linguistiquement liés. Plusieurs exemples sont présentés dans la table II.1[3].

mot	**stem**
déménageuse déménagement déménager déménageur	déménag
porter porteur portatif port portuaire	port

TAB. II.1 – Exemples de stems

Le "stemming" est également réalisé sur la requête avant d'effectuer la correspondance entre la représentation de la requête et les représentations des documents. Le moteur de recherche sélectionne un ensemble de documents candidats, puis ces documents sont ordonnés et utilisés pour enrichir l'historique de questions. Ces historiques peuvent être utiles par la suite pour enrichir la représentation d'une nouvelle requête qui semble être proche de cet historique. Les mots clés des réponses satisfaisantes aux questions similaires peuvent être utilisés pour enrichir la représentation de la requête actuelle avec de nouveaux termes. Ce processus, guidé par l'utilisateur, qui exprime la satisfaction à une question, est appelé retour de pertinence ou "relevance feed-back".

Les approches en recherche documentaire peuvent être classées en deux catégories. La première est basée sur les correspondances entre vecteurs de paramètres sur les statistiques des mots. La seconde s'appuie sur les représentations conceptuelles, qui peuvent être basées sur la sélection de termes. Dans ce cas, les termes définissent le document et sont donnés par un expert humain. La correspondance entre les mots et la représentation conceptuelle est obtenue par une source de connaissances qui est souvent un thesaurus[4] de sens et de synonymes des mots.

[2]Liste de mots. On trouve également les termes *anti-dictionnaire* ou *anti-lexique*
[3]déménageuse : Mot Suisse qui signifie camion de déménagement
[4]Recueil documentaire alphabétique de termes servant de descripteurs pour analyser un corpus

Des recherches récentes sur l'utilisation des représentations sémantiques pour la recherche documentaire ont mis en évidence les problèmes causés par l'utilisation des mots pour représenter le contenu des documents ([Krovetz 1995]). Deux problèmes émergent alors : l'ambiguïté des mots, et le fait qu'un document pertinent pour une question ne contiendra pas forcément les mêmes mots que ceux de la question posée, même s'il indique le même concept. Il est donc important de prendre en compte les homonymies[5], les polysémies[6], les synonymies, les hyperonymies[7], les hyponymies[8] ... S'il était possible d'obtenir le sens de chacun des mots, un thesaurus pourrait suggérer les mots dont le sens est relatif et ces mots pourraient être utilisés pour enrichir la question. Mais ceci cause des problèmes car il est courant qu'un thesaurus soit réalisé manuellement pour une application et ne puisse pas être réutilisé pour d'autres.

Notre étude porte sur les systèmes basés sur les statistiques des mots. Les modèles probabilistes conventionnels sont basés sur la théorie de décision bayésienne. Les réponses candidates sont obtenues à partir des données contenues dans la question. Ces données peuvent aussi être augmentées par retour de pertinence. Il existe trois choix fondamentaux dans la conception des systèmes de recherche documentaire ([Wong et Yao 1995]) et selon la manière dont ces problèmes sont traités, différents modèles de recherche peuvent être considérés : le dispositif pour la représentation des documents, la formulation de la requête, et la construction d'une fonction d'ordonnancement qui exprime le degré de pertinence d'un document pour une requête. Nous nous intéressons plus particulièrement au dispositif pour représenter les documents car celui-ci implique une catégorisation hiérarchique des documents qui peut inclure une classification et segmentation thématique, problèmes que nous abordons aux chapitres III et V.

1.1 Modèles probabilistes

Représenter les documents sous forme d'un espace vectoriel est une technique largement utilisée en recherche documentaire. On trouve ce type de représentation des documents dans [Salton et al. 1975]. On considère une collection de documents $D = \{d_1, d_2, d_3, d_4\}$, un dispositif de représentation consistant en un ensemble de termes $T = \{t_1, t_2, t_3\}$ et une matrice termes/documents dans laquelle l'élément de la i-ème ligne et la j-ème colonne représente le nombre de fois où le terme t_j a été observé dans le document d_i :

	t_1	t_2	t_3
d_1	2	0	1
d_2	1	0	1
d_3	0	1	3
d_4	2	1	0

Il est alors possible de calculer une similarité entre documents $s(D_i, D_j)$ et de les regrouper en classes selon différents critères. Plusieurs solutions sont possibles pour affecter un poids aux termes des documents ([Salton et Buckley 1988]). Dans les modèles de type *booléens*, une question est exprimée par une formule logique de termes et un document est proposé seulement s'il satisfait logiquement la requête. Une telle règle stricte est plus appropriée aux bases de

[5]Homonymes : le *port* de l'uniforme/le *port* de commerce, un *livre* intéressant/une *livre* de beurre
[6]Polysémies : la *construction* du pont/une *construction* moderne, la *clé* de la porte/la *clé* du problème
[7]Hyperonymies : *insecte* est l'hyperonyme de *mouche*, *abeille* ou *fourmi*
[8]Hyponymies : *mouche*, *abeille* et *fourmi* sont des hyponymes de *insecte*

données qu'à la recherche documentaire. Une correspondance moins stricte a été proposée en utilisant la logique floue. Une présentation du problème est dans [Van Rijsbergen 1986].

Les modèles probabilistes modernes sont basés sur la notion de *probability of relevance* ([**?**]). Les poids des termes dans les questions et les documents sont également vus sous forme de probabilités. Dans ces modèles, on définit un espace conceptuel U dont les éléments sont des concepts élémentaires. On définit aussi $P(R \mid q, d)$ comme étant la probabilité de pertinence R étant donnés une question q et un document d. La fonction probabiliste P est définie sur U telle que :

- $P(d)$, pour un document d, est le degré de couverture dans U des concepts de d,
- $P(q, d)$ représente le degré pour lequel U est couvert par les connaissances communes du document d et de la question q.

Plusieurs solutions sont possibles pour évaluer $P(R \mid d, q)$. Dans les systèmes de recherche par probabilités jointes, on considère l'approximation suivante :

$$P(R \mid q, d) \approx P(q, d)$$

Les fonctions booléennes $\tau(t, d)$, $\tau(t, q)$ renvoient *vrai* ssi le second argument contient le terme t, et on définit :

$$P(q, d) = \sum_t P(t)\tau(t, d)\tau(t, q)$$

Si $P(t) = k$ alors :

$$P(q, d) = k \parallel d \cap q \parallel$$

où la probabilité est proportionnelle à l'union des termes communs à d et q. Si maintenant $P(t)$ est remplacée par l'inverse de la fréquence d'un document (IDF) :

$$idf(t) = -\log \frac{n(t)}{N}$$

où $n(t)$ est le nombre de documents dans lesquels t est présent, N est le nombre total de documents, et les fonctions booléennes $\tau(t, d)$, $\tau(t, q)$ sont remplacées par les fréquences des occurrences du terme t dans un document et une question q, alors, le score suivant est obtenu :

$$P(q, d) = \sum_t P(t)tf_d(t)tf_q(t)$$

D'autres possibilités peuvent être utilisées pour représenter $P(q, d)$. [Turtle et Croft 1990] proposent un modèle de recherche basé sur les réseaux d'inférence bayésiens. Ce réseau est un graphe de dépendances acyclique dans lequel les nœuds représentent des variables propositionnelles ou des constantes et les arcs représentent des relations de dépendance entre propositions. Si une proposition, représentée par un nœud p, cause ou implique une proposition représentée par un nœud q, ils établissent un arc direct entre p et q. Le nœud q contient une matrice de liens qui spécifie $P(q \mid p)$ pour toutes les valeurs possibles des deux variables. Quand un nœud a plusieurs parents, la matrice de liens spécifie l'ensemble de ses dépendances avec l'ensemble de ses parents (π_q). Le réseau peut alors être utilisé pour évaluer une probabilité ou un degré d'association entre nœuds. De même, dans [Greiff et Croft 1999], on fait référence aux réseaux bayésiens qui utilisent une distribution de probabilités jointes. Les nœuds du réseau correspondent à des variables aléatoires qui peuvent prendre les valeurs booléennes et la topologie du réseau peut être vue comme un ensemble de relations conditionnelles d'indépendance entre variables.

Un des problèmes inhérents aux méthodes citées dans cette section est la masse importante de données nécessaires à la représentation des documents. Dans le cas des réseaux bayésiens, la topologie du réseau peut être réduite, mais reste importante. Dans le cas des systèmes par représentation vectorielle, une solution possible pour réduire l'espace de représentation est d'appliquer une décomposition SVD[9].

1.2 Evaluation des systèmes de recherche documentaire

Les systèmes de recherche documentaire sont évalués sur leur temps de réponse et sur la complexité associée à leurs algorithmes et leurs structures de données, ainsi qu'avec les valeurs de rappel et précision qui sont introduits dans cette section pour les systèmes probabilistes.

Etant données les propositions E et H, l'implication $E \rightarrow H$ indique que E supporte H, i. e. H peut être vu comme une hypothèse et E comme l'évidence qui l'appuie. La double implication $E \leftrightarrow H$ signifie que E et H se supportent mutuellement. On donne :

$$\psi(E \rightarrow H) = P(H|E) = \frac{P(H \cap E)}{P(E)} \quad \text{et} \quad \psi(E \leftrightarrow H) = \frac{P(H \cap E)}{P(E \cup H)} \tag{II.1}$$

Mais en pratique, d'après [Wong et Yao 1995], on connaît peu de choses sur les concepts élémentaires de U. Par contre, on peut avoir des informations sur K, un sous-espace de U. Par exemple, K peut être l'union des termes (mots-clés) utilisés comme indices (connaissance représentative) pour un document. Dans ce cas :

$$\psi(E \rightarrow H|K) = P(H|E \cup K) = \frac{P(H \cap E \cap K)}{P(E \cap K)} \tag{II.2}$$

Une mesure de la précision, telle que la proposition d représente la connaissance d'un document et la proposition q représente l'information demandée, est la suivante :

$$precision = \psi(d \rightarrow q) = P(q|d) = \frac{P(q \cap d)}{P(d)} \tag{II.3}$$

qui est la probabilité que, étant donné un document proposé, celui-ci soit pertinent pour la question.

Une valeur de rappel peut être définie comme suit :

$$rappel = \psi(q \rightarrow d) = P(d|q) = \frac{P(d \cap q)}{P(q)} = \frac{P(d)}{P(q)}.\psi(d \rightarrow q) \tag{II.4}$$

qui est la probabilité que, étant donné une question, un document pertinent soit proposé.

Les listes de documents rapportées par un système de recherche sont souvent très longues. L'ordonnancement des documents par *ordre de précision* facilite l'exploration mais ne permet pas de détecter facilement les thématiques abordées par les textes. De plus, les sujets abordés par ces documents sont multiples et ne correspondent pas toujours à ce qu'attendait l'utilisateur. La classification thématique des documents trouvés répond à ces problèmes. Par ailleurs, la segmentation des documents est également une manière de répondre à ces problèmes, en ce sens qu'elle ne présentera que la partie pertinente du document, et non la totalité. Les sections suivantes abordent les problèmes de classification et de segmentation, notamment selon un critère thématique.

[9]Singular Value Decomposition

2 Classification

La classification est le regroupement d'entités textuelles en classes et l'assignement d'une ou plusieurs de ces classes à un texte donné. Les classes peuvent ensuite être utilisées comme référent pour les documents. Différentes méthodes sont possibles pour réaliser la classification. Nous aborderons les méthodes basées sur les règles d'induction et les méthodes statistiques comme les systèmes de classification bayésiens, les arbres de classification et plus particulièrement les systèmes à base de mots clés.

2.1 Méthodes à base de règles

Dans [Apté et al. 1994], une méthode basée sur les règles d'induction est utilisée pour assigner des catégories prédéterminées aux textes. Un corpus de textes est étiqueté pour réaliser la phase d'apprentissage des règles qui permettront de classifier de nouveaux textes inconnus. Ces règles sont basées sur l'occurrence des mots et des expressions particulières dans les textes. Par exemple, un texte contenant l'expression "running back" serait immédiatement assignée à la catégorie "football" ; on peut dire de même des éléments contenant les deux mots "award" et "player". Des méthodes heuristiques sont employées pour trouver les règles simples qui séparent correctement toutes les classes. Le fonctionnement général du système est indiqué dans la figure II.3. Les auteurs ont appliqué leur méthode sur des articles diffusés en langues

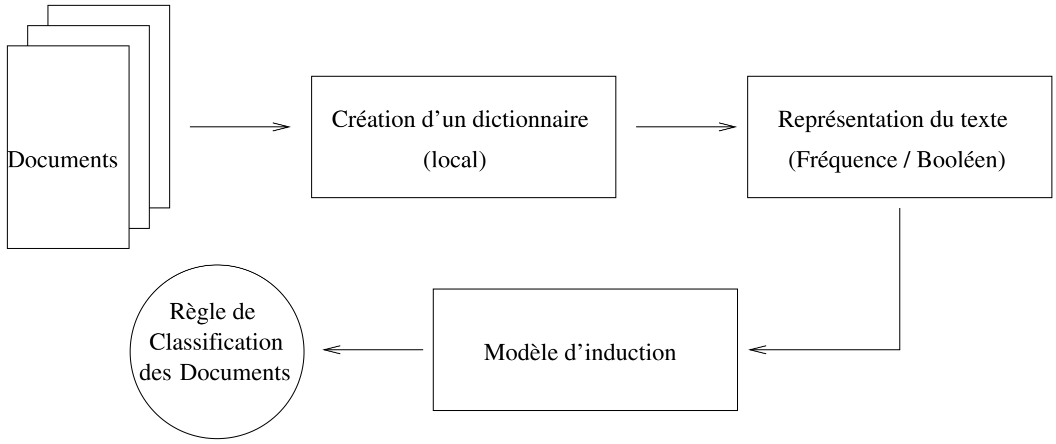


FIG. II.3 – Architecture de l'apprentissage automatique à base de règles d'induction

allemande et anglaise. Ils calculent des paramètres d'évaluation tels que le pourcentage des documents totaux qui sont correctement classifiés pour une catégorie donnée, et constatent que leur système fonctionne avec un degré élevé d'exactitude. De plus, aucune différence n'a été détectée entre les collections de langues anglaises et allemandes, laissant supposer que la méthode peut fonctionner sur un nombre important de langues.

Dans [Junker et Abecker 1997], on propose d'utiliser un ensemble d'exemples de documents préclassés par la représentation d'un langage. Les règles d'induction sont construites de telle sorte que la tête correspond à une classe et le corps spécifie des tests à réaliser sur les documents. Le système, tel qu'il est décrit dans la figure II.4, consiste en un module qui génère des hypothèses de règles formulées dans une hypothèse de langage. Enfin, les règles sont associées aux synsets de WordNet.

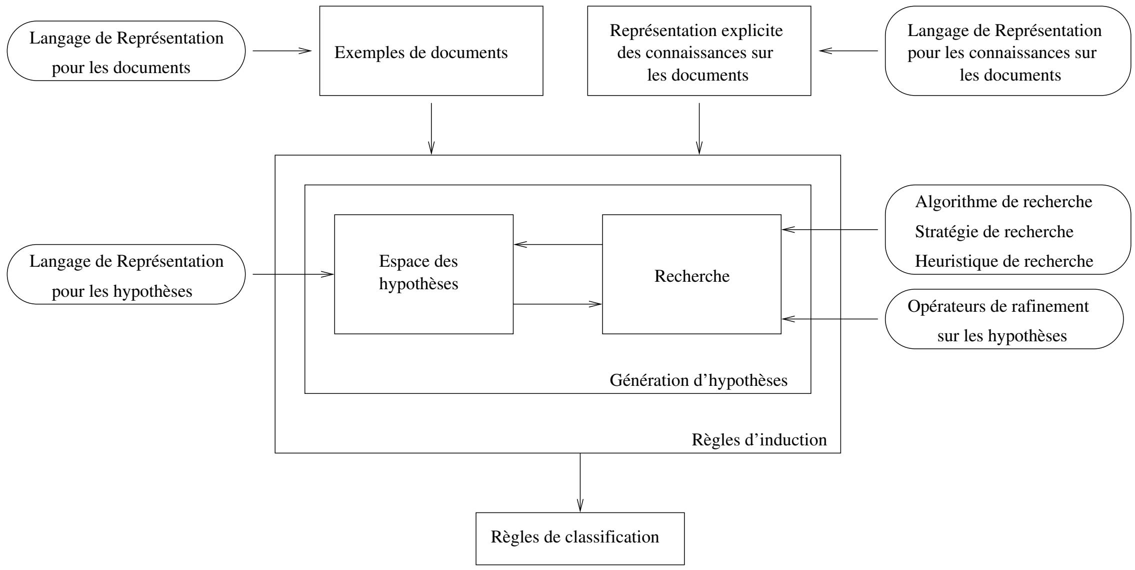


FIG. II.4 – Paramètres qui influencent la recherche des règles

Dans [Stuart et al. 1990], on décrit une approche d'apprentissage automatique pour la construction automatique d'arbres de classification pour la recherche documentaire. Leur système de recherche documentaire (CART)[10] prend en entrée une collection d'exemples de données d'apprentissage dont on a assigné une classe et un vecteur de caractéristiques. La sortie de CART est un arbre de décision binaire qui peut être utilisé pour classifier des observations dont la classe n'est pas connue. Chacun des articles qui est utilisé pour l'apprentissage est étiqueté selon qu'il appartient ou non à un thème (i. e. l'etiquette est "thème" ou "¬thème"). Par exemple, 82 mots ont été sélectionnés comme étant pertinents pour le thème *terrorisme*. Le corpus d'apprentissage pour ce thème a été généré en comptant le nombre d'occurrences de chacun de ces mots dans chacun des articles. Les arbres de classification sont, par exemple, comme suit :

Si *l'article contient le mot "bomb"*
Alors *Si l'article contient le mot "injure" ou "kill"*
Alors le thème est ***"terrorism"***
Sinon le thème est ***"¬terrorism"***
Sinon *Si l'article contient le mot "kidnapping"*
Alors le thème est ***"terrorism"***
Sinon le thème est ***"¬terrorism"***

Par la suite, des sous-concepts ont été introduits. Deux lecteurs indépendants ont indiqué quels sont les sous-concepts pertinents (parmi les 18 possibles) pour chacun des documents. Ainsi, un nouvel apprentissage est réalisé en générant des arbres qui donnent une valeur booléenne selon que le sous-concept est présent ou non dans chaque article. On obtient un arbre, par exemple, tel que :

[10] Classification and Regression Trees

Si	*l'article contient le sous-concept "bombing"*
Alors	*Si l'article contient les sous-concepts "explosion" ou "killing"*
	*Alors le thème est **"terrorism"***
	*Sinon le thème est **"¬terrorism"***
Sinon	*Si l'article contient le sous-concept "kidnap-event" et "named-terrorist"*
	*Alors le thème est **"terrorism"***
	*Sinon le thème est **"¬terrorism"***

Le problème dans ce type de méthodes est qu'elles nécessitent un pré-classement manuel afin de construire le référent. Dans [Bellot et El-Bèze 2000], on propose une méthode alternative à la classification hiérarchique et aux méthodes de partitionnement. Pour construire un arbre de décision, on doit définir un ensemble de questions, une règle pour déterminer la meilleure question à poser aux textes d'un nœud et un critère d'arrêt déterminant l'ensemble des feuilles de l'arbre. Dans ce cas, il est possible de considérer les phrases des textes comme étant les individus contenus dans les nœuds de l'arbre. La racine de l'arbre contient ainsi toutes les phrases des documents trouvés par le système de recherche documentaire pour une requête. Le but étant de regrouper les individus proches les uns des autres, on fait en sorte de placer dans la même feuille de l'arbre ceux qui ont en commun le plus grand nombre possible de mots. A chaque nœud de l'arbre, il s'agit donc de calculer quel est le mot x qui permet de subdiviser au mieux les individus qu'il contient. Une question Q_x de la forme *"les individus contiennent-ils le mot x ?"* est posée à chaque individu. Elle subdivise un nœud N en deux fils N_{OUI} et N_{NON}.

Des questions "doubles" de la forme *"les individus contiennent-ils les mots x et y ?"* peuvent également être posées, en imposant que l'un des deux termes x ou y soit issu de la requête. Le critère de qualité d'une classe dépend de la pertinence des individus qu'elle contient. La pertinence est calculée suivant les probabilités associées à chaque classe. L'entropie permet de mesurer l'homogénéité des phrases dans les nœuds de l'arbre en fonction de leurs niveaux respectifs de pertinence par rapport à la requête. Pour cela, sont calculées pour chaque question N : l'entropie H liée au nœud considéré N et l'entropie moyenne des nœuds N_{OUI} et N_{NON} résultant de la subdivision entraînée par Q. La question choisie est celle à laquelle correspond la plus grande baisse d'entropie.

2.2 Systèmes de classification bayésiens

Dans [Lewis et Gale 1994], on propose une conjonction avec un système de classification bayésien. Il consiste en l'apprentissage d'un classificateur avec un nombre limité de données annotées manuellement et utilise ces connaissances pour classer les nouveaux documents. Par la suite, le système de classification est réappris avec les nouvelles données ainsi formées. Cette itération est répétée avec très peu de nouvelles données entrées à chaque étape, jusqu'à l'insertion de toutes les données. Ce système probabiliste de classification de textes considère un ensemble $W = (w_1 \cdots w_I)$ de caractéristiques extraites à partir des documents et affecte le document à une classe C_j telle que la probabilité suivante soit maximale :

$$P(C_j \mid W) = \frac{P(W \mid C_j).P(C_j)}{\sum_{k=1}^{K} P(W \mid C_k).P(C_k)}$$

Les auteurs proposent de ne considérer que 2 classes possibles, telles que $C_1 = C$ et $C_2 = \widetilde{C}$, avec $P(\widetilde{C}) = 1 - P(C)$. Dans ce cas, en évaluant les probabilités avec un modèle de langage *unigramme*, on obtient la valeur de *odd ratio* telle que :

$$\frac{P(C \mid W)}{P(\widetilde{C} \mid W)} = \frac{P(C)}{P(\widetilde{C})} \prod_{i=1}^{I} \frac{P(w_i \mid C)}{P(w_i \mid \widetilde{C})}$$

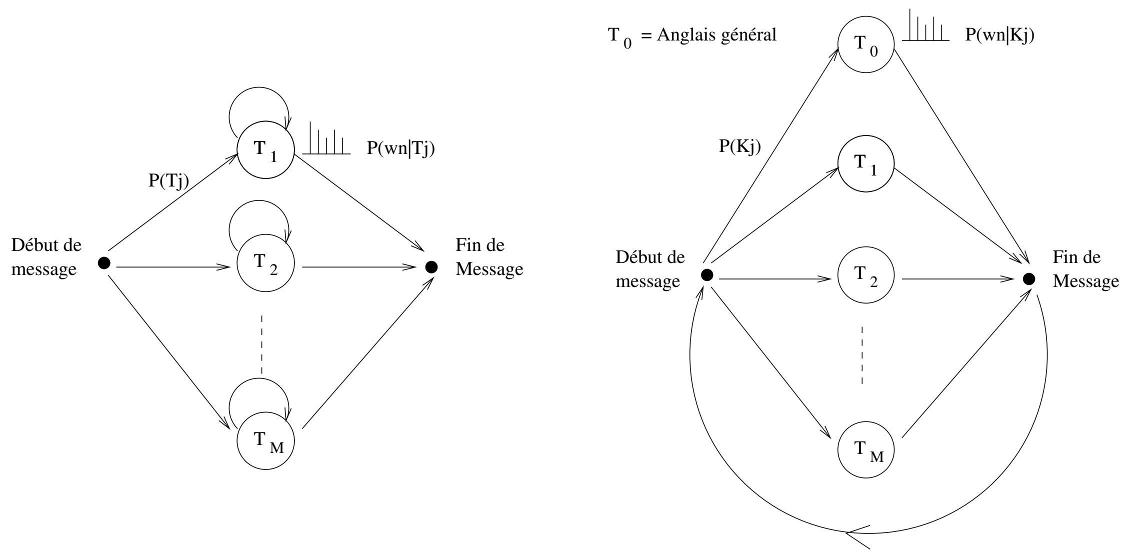
La régression logistique[11] est une technique généralement utilisée pour combiner plusieurs valeurs afin d'estimer les probabilités (a et b servent à amortir les ratios de probabilité résultant des violations de l'indépendance) :

$$P(C \mid w) = \frac{\exp(a + b \sum_{i=1}^{I} \log \frac{P(w_i|C)}{P(w_i|\widetilde{C})})}{1 + \exp(a + b \sum_{i=1}^{I} \log \frac{P(w_i|C)}{P(w_i|\widetilde{C})})}$$

où on peut définir $a = \log \frac{P(C)}{P(\widetilde{C})}$.

2.3 Modèles de Markov Cachés thématiques

[Imai et al. 1997] utilisent un nouveau modèle thématique composé d'un simple HMM, où chaque état du HMM représente un thème. La figure suivante montre un modèle conventionnel et les modèles thématiques qu'ils proposent.



Leur particularité est qu'ils pensent que l'on peut passer d'un thème à l'autre à chaque nouveau mot, ils considèrent également qu'une histoire peut être composée de plusieurs thèmes. De plus, ils effectuent leurs expériences avec un très grand nombre de thèmes différents portant sur des sujets précis. C'est pourquoi, ils affectent 4 étiquettes thématiques à chaque message.

[11] Soit Z une variable qualitative à 2 modalités : 1 ou 0, succès ou échec. La régression logistique s'interprète comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio", c'est-à-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.

2.4 Extraction et sélection des mots clés

Dans [Ohtsuki et al. 1998], on propose des méthodes pour extraire les mots thématiques d'un corpus. Plutôt que la classification d'un document dans des thèmes, il est possible de définir le thème d'un document avec un ensemble de mots thématiques. Ceux-ci sont obtenus en identifiant un ensemble global de mots thématiques faits de tous les noms et verbes des titres des documents. Etant donné un document, les 5 mots thématiques qui sont les plus proches du contenu du document sont sélectionnés. La proximité est mesurée avec une distance et un critère du χ^2.

La distance est définie en utilisant l'information mutuelle, comme suit :

$$d(w_i, t_j) = P(w_i, t_j) \log \frac{P(w_i, t_j)}{P(w_i)P(t_j)} \tag{II.5}$$

où t_j est un mot thématique et w_i est un mot du document. Le critère du χ^2 est basé sur la mesure suivante :

$$\chi^2_{ij} = \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \tag{II.6}$$

où f_{ij} est la fréquence de w_i apparaissant dans un document avec t_j et :

$$F_{ij} = \frac{\sum_{l=1}^{M} f_{il}}{\sum_{k=1}^{|V|} \sum_{l=1}^{M} f_{kl}} \tag{II.7}$$

où M est le nombre de mots thématiques distincts.

Le critère du χ^2 semble être meilleur que le critère de distance dans une expérience dans laquelle, étant donné un article, les utilisateurs proposent des mots thématiques qui sont comparés avec ceux obtenus par la procédure automatique. Un mot obtenu est correct s'il appartient à la disjonction de l'ensemble de mots thématiques proposés par un groupe d'experts. Les expériences ont étés réalisées sur des documents de parole dans lesquels les mots sont le résultat d'un processus de reconnaissance. Le problème dans ce type d'approche est qu'elle requiert la connaissance d'experts et donc une validation manuelle.

Dans [Wright et al. 1995], les mots clés sont choisis automatiquement, à partir d'un corpus d'apprentissage, sur la base de leur contribution relative à discriminer les thèmes. Ils sont sélectionnés sur le critère suivant :

$$P(w_k|T) \log \frac{P(w_k|T)}{P(w_k|\overline{T})} \tag{II.8}$$

Les distributions statistiques des mots clés sont obtenues à partir de données d'apprentissage lissées avec une distribution de Poisson.

Dans [Yamashita et al. 1998], la sélection des mots clés s'effectue par la mesure de l'information mutuelle entre un mot et un thème, $I(T;W)$, où on prendra en compte la contribution qu'un mot w peut apporter pour l'identification de T. Par ailleurs, les mots inférieurs à 6 phonèmes sont omis des mots clés. L'identification thématique d'informations télévisées est basée sur des techniques de repérage des mots clés à partir du signal acoustique. Un thème est identifié pour une fenêtre de M secondes, appelée fenêtre d'analyse thématique (TAW - *Topic Analysis Window).* Le repérage des mots clés génère une séquence de mots clés w. Le thème le plus probable est déterminé en calculant un score pour chaque thème en fonction des probabilités de la séquence des mots clés.

Le score $F(T_i)$ du i-ème thème T_i est défini tel que :

$$F(T_i) = \frac{\log P(T_i \mid w) + k.R(w)}{N} - B(T_i)$$

où $R(w)$ est le score total de reconnaissance des mots clés dans leur phase de repérage, k est un facteur de pondération et N est le nombre de mots clés détectés dans une fenêtre TAW. $B(T_i)$ dénote un "biais" pour le i-ème thème dans son ensemble de mots clés.

2.5 Similarité du contenu des messages

Dans [Carlson 1996], on propose une approche statistique afin de regrouper des enregistrements de parole. La méthode s'appuie sur la similarité des contenus thématiques des messages. L'algorithme proposé est basé sur des modèles de langage n-grammes et du clustering dans les arbres hiérarchiques.

L'algorithme de clustering qui est proposé est le suivant, où MDS signifie échelonage multi dimentionnel, et ROC, Receiver Operator Curve :

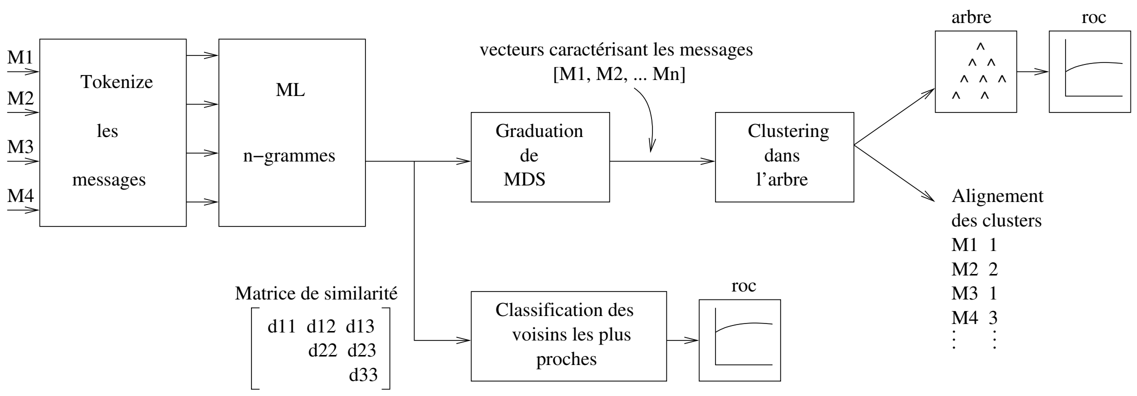


Fig. II.5 – Vue d'ensemble de l'algorithme de clustering

Pour chaque message, un modèle est construit et la distance entre deux messages est la somme des logarithmes des ratios de probabilités de chacun des messages par rapport à l'autre modèle :

$$d(M_i, M_j) = \log \frac{P(M_i|L_i)}{P(M_i|L_j)} + \log \frac{P(M_j|L_j)}{P(M_j|L_i)} \tag{II.9}$$

3 Segmentation

Rapporter de longs textes en réponse à la question d'un utilisateur n'est pas efficace car l'utilisateur doit examiner beaucoup de données avant d'atteindre ce qu'il cherche. L'isolation de certaines parties pertinentes d'un texte permet de mieux le positionner dans la liste des réponses et de fournir rapidement à l'utilisateur une indication sur la localisation précise de ce qu'il cherche. Plusieurs méthodes de segmentation vont être maintenant présentées.

3.1 Classification hiérarchique

Dans [Yaari 1997], la segmentation est réalisée grâce à l'utilisation d'un algorithme de classification hiérarchique. La méthode utilise les paragraphes comme segments de base pour identifier la structure hiérarchique du discours. Il évalue une similarité lexicale comme test de proximité entre les paragraphes. Tant qu'il reste des segments non classés, les deux segments adjacents les plus proches sont regroupés hiérarchiquement. Les seules distances calculées pour construire cette hiérarchie sont les distances séparant un paragraphe de celui qui le suit. Les frontières thématiques sont déduites de la hiérarchie suivant les profondeurs respectives des segments adjacents dans le texte.

Dans [Bellot et El-Bèze 2000], l'application de l'algorithme présenté permet de réunir les phrases issues des documents trouvés en réponse à une requête en un certain nombre de classes. Cette classification permet souvent un regroupement thématique correct. Il est donc possible de supposer que deux phrases issues de la même classe partagent la même thématique. A l'opposé, si deux phrases se trouvent dans deux classes différentes, on suppose qu'elles abordent des sujets différents. Il est décidé dans ce cas d'apposer entre elles une marque de frontière thématique et l'on procède de même pour l'ensemble des documents rapportés. Une segmentation est ainsi directement déduite de la classification.

3.2 Programmation dynamique

Dans [Heinonen 1998], l'algorithme de segmentation consiste à unifier les paragraphes successifs d'un texte (s'il y a lieu ...) en utilisant la programmation dynamique. Ils commencent avec la première borne possible et évaluent un coût pour cette limite possible comme si le premier paragraphe consistait en un seul segment. Par la suite, ils traitent la deuxième borne et lui attribue la valeur minimale parmi les deux possibilitées suivantes : le coût des deux premiers paragraphes comme s'ils constituaient un seul fragment de texte et le coût du deuxième paragraphe comme constituant un fragment de texte séparé. Dans les étapes suivantes, l'évaluation s'effectue paragraphe par paragraphe, et on considère toutes les possibilités de localisation des points de césure. La procédure continue ainsi jusqu'à la fin du texte, et, au final, on génère une liste de césures indiquant la segmentation du texte.

3.3 Cohésion Lexicale

La cohésion lexicale entre deux discours est un indicateur de leur cohérence textuelle et elle s'évalue quand les segments contiennent des mots pour lesquels il existe une relation de similitude ou une relation sémantique. On peut donc dire qu'il existe une segmentation possible entre deux segments si leur cohésion lexicale, évaluée par des mesures de similarité, atteint un certain seuil.

Dans [Ferret 1998] et [Ferret et Grau 1998], on propose une segmentation thématique des textes reposant sur la cohésion lexicale renfermée par un vaste réseau de cooccurrences lexicales. C'est une méthode fondée sur l'analyse de la cohésion thématique en tout point d'un texte. Il prend la forme canonique des mots pleins (noms, verbes, adjectifs) en éliminant les noms propres, abréviations et certains verbes. Il apprend un réseau de cooccurrences lexicales dont le fonctionnement est assimilable à celui d'un cache de 20 mots (c'est une fenêtre qui est apprise). Pour procéder à la re-découverte des frontières de textes, il calcule la cohésion en faisant glisser une fenêtre sur le texte. S'il y a une forte cohésion entre les données d'une fenêtre à un instant donné, c'est que la fenêtre est de même thème, sinon, il y a des thèmes différents. La cohésion résulte de la combinaison, pour chaque mot impliqué, d'un poids dépendant de son contexte d'apparition et d'une mesure générale de sa capacité discriminante sur le plan thématique. Leur méthode est adaptée au suivi des évolutions finies des thèmes d'un texte, et non à la mise en évidence des changements plus radicaux.

Dans [Jobbins et Evett 1998], la segmentation est réalisée en comparant des fenêtres adjacentes de texte pour déterminer leur similarité lexicale. Les relations de cohésion lexicales, réitérations et collocation sont utilisées pour identifier les mots relatifs les uns aux autres. Ces relations sont automatiquement localisées avec une combinaison de plusieurs caractéristiques linguistiques. La cohésion lexicale s'exprime à travers le vocabulaire utilisé dans le texte et les relations sémantiques entre ceux-ci. La cohésion lexicale est divisée en trois classes, dont la première est ignorée :

1. les noms généraux,
2. les réitérations,
3. les collocations.

La *réitération* est sous-divisée en plusieurs effets de cohésion : la répétition des mots, les synonymes et les mots généraux. Une *collocation* est une combinaison de mots "prédisposée", typiquement les paires de mots qui tendent à co-occurrer régulièrement. D'autres relations entre les mots sont également prises en compte, avec l'utilisation d'un thésaurus. Ce sont des relations qui quantifient la relation d'ordre sémantique entre les mots. Ces relations étant établies, leur méthode de segmentation consiste à comparer des fenêtres adjacentes et détermine leur similarité lexicale. La taille de fenêtre qui donne les meilleurs résultats est de 3 phrases. La similarité lexicale est calculée pour chaque fenêtre en effectuant une comparaison basée sur la proportion de mots en relation et leur score de normalisation. Les répétitions de mots sont identifiées entre les mots identiques et à partir des mots qui ont les mêmes stems. Les collocations sont localisées en observant les paires de mots dans un lexique de collocations. Les relations sémantiques sont calculées entre paires de mots selon leur localisation dans le texte.

3.4 Modèles de Markov Cachés

[Yamron et al. 1997] présentent une approche aux problèmes de la segmentation de texte et de la détection d'événements, par l'utilisation des modèles de Markov cachés (HMM) et des techniques de clustering. Pour ce faire, ils proposent de considérer un flot de texte non segmenté comme étant composé d'une série de sujets de la même manière qu'un flot de parole consiste en une série de phonèmes. Pour la tâche de segmentation du texte, les histoires sont considérées comme des instances de thèmes, et chacun d'entre eux est caractérisé par un modèle statistique sur l'espace des chaînes de mots. De plus, les probabilités sur les transitions entre les thèmes sont évaluées. Dans ce cas, étant donné un flot de texte, une probabilité peut être attribuée à chaque hypothèse particulière sur la séquence et la segmentation en thèmes s'obtient en appliquant l'algorithme suivant :

1. La transition de l'état de départ au premier thème, accumuler une probabilité de transition.
2. Rester dans le thème un certain nombre de mots ou phrases, et, étant donné le thème courant, accumuler une probabilité de boucle sur lui-même et une probabilité de modèle de langage pour chacun.
3. Transition vers un nouveau thème, accumuler la probabilité de transition. Retour en 2.

La recherche de la meilleure hypothèse et la segmentation correspondante peut être réalisée en utilisant les techniques HMM standards, et les techniques connues en reconnaissance de la parole.

Dans [van Mulbregt et al. 1998], l'objectif est la transcription automatique de messages de radio : segmentation des messages en histoires et recherche de ces histoires relatives à un thème spécifique. La méthode effectue la recherche de la meilleure hypothèse de segmentation par un HMM. Le segmenteur associe un ML à chaque thème, et affecte une probabilité de transitions entre les thèmes (qui implique une "durée" pour chaque thème). Cette segmentation produit un label pour chaque segment. Par la suite, la méthode de clustering en "multi-pass k-means algorithm" nécessite plusieurs passes sur le texte. Les résultats sont très différents selon les corpora car leur HMM tient compte de la longueur des segments. Dans leur système, c'est le segmenteur qui affecte un label thématique aux segments de textes. Mais leurs thèmes ne correspondent pas à des thèmes comme peut l'entendre un humain, ce sont des "clusters".

4 Classification et segmentation en recherche documentaire de documents de parole

Le but de la recherche documentaire adaptée aux documents de parole est de placer dans des catégories une expression de la parole selon une notion prédéfinie de classes thématiques ou de messages. La classification de documents de parole est motivée par les systèmes de stockage et de transmission des messages. Une autre application peut être la classification d'appels téléphoniques, pour diriger les appels de clients vers les zones de services appropriées.

[McDonough et Gish 1994] proposent que la tâche d'identification thématique soit partagée en 3 sous-problèmes :

1. détection d'événements,
2. sélection de mots clés,
3. modélisation thématique.

La première tâche consiste à extraire les caractéristiques et les événements pertinents d'un message qui pourront, par la suite, être utilisés pour la modélisation thématique et l'identification. Les événements qui sont considérés sont le nombre d'occurrences de chacun des membres d'un ensemble de mots clés ; chaque détection correcte de mot clé produit une information concernant le thème de la discussion abordée dans le message. Leurs recherches précédentes ont montré que la classification thématique est meilleure lorsque l'on considère uniquement les mots clés plutôt que tous les mots du message. Deux méthodes sont utilisées pour la sélection des mots clés : l'une basée sur la distance de Kullback-Leibler symétrique ([Cover et Thomas 1991]), l'autre sur la mesure d'information mutuelle. Le repérage de mots est obtenu en considérant la probabilité $P(w,t)$ qu'un mot w soit terminé au temps t. Un mot-clé w est considéré comme étant détecté si son score obtient un maximum local au temps t. Ce score est évalué en considérant α et β pour l'état de fin d'un mot w, tel que :

$$P(w,t) = \frac{\alpha(e_w,t)\beta(e_w,t)}{\sum_{allstates}\alpha(s,t)\beta(s,t)} \tag{II.10}$$

où α est la probabilité avant, β est la probabilité arrière de l'algorithme de Baum-Welch, et e_w est l'état final du modèle de Markov du mot w. La somme de $P(w,t)$ pour tous les meilleurs mots w est considérée comme étant égale au nombre de fois où w a été observé dans le message.

Le repérage des mots clés est également utilisé dans un système décrit par [Roses et al. 1991]. Ce système génère des hypothèses de mots basées sur ce qui est reconnu du message de parole. Tous les types de messages utilisent le même vocabulaire de mots clés. Un message M est perçu comme un ensemble de mots indépendants, dont il existe un sous-ensemble parmi les K mots qui forment le vocabulaire de classification du message. Pour chacun des mots du vocabulaire, il existe un classifieur de messages binaire qui est activé par l'occurrence d'un mot du vocabulaire dans le message entré, tel que :

$$\log P(C_j|M) = \sum_{k\in V}\log\frac{P(C_j,w_k)}{P(C_j)P(w_k)} + \log P(C_j) \tag{II.11}$$

où M est le message en entrée et C_j l'ensemble des classes possibles. On peut considérer que la classe C_j représente un thème T_j, et que le message M est la suite de mots W_1^n. La confiance à accorder aux mots labélisés est pondérée avant de déclencher le classifieur. Ces pondérations sont estimées en utilisant des algorithmes de propagation arrière pour minimiser les erreurs de classification. Ces pondérations sont :

$$V_{k,j} = \log\frac{P(w_k|T_j)}{P(w_k)} + \log P(T_j) \tag{II.12}$$

Si V est le vocabulaire des mot-clés, alors le score pour le thème T_j est :

$$s_j = \sum_{k\in V} Vi_{k,j}a_k \tag{II.13}$$

$$a_k = 1$$

si et seulement si w_k est labélisé Les coefficients peuvent aussi être un score pondéré avec le label du mot correspondant. Cette approche ne tient pas compte du nombre de fois où un mots est observé dans un document.

Il est possible aussi d'utiliser directement les sous-chaînes de phonèmes pour réaliser la classification et la segmentation de textes de parole. Les modèles de mots clés sont obtenus par caractérisation des modèles de phonèmes selon la séquence de phonèmes de la prononciation des mots clés. Tous les autres mots sont représentés par un modèle de remplissage ("filler"). Ce dernier n'est fait que de tous les modèles de phonèmes mis en parallèle avec les silences.

Différents types de modèles de phonèmes peuvent être utilisés pour la recherche documentaire comme décrit dans [Foote et al. 1997]. En fait, les modèles de mots clés peuvent être faits avec des modèles monophones (indépendants du contexte), des modèles de phonèmes ou biphones, ou encore triphones (le phonème dans le contexte). Le modèle de remplissage est toujours obtenu avec des modèles monophones. De plus, les modèles peuvent être dépendants du locuteur (SD), indépendant du locuteur (SI) ou adaptés au locuteur (SA). Dans [Foote et al. 1997] on remarque que l'utilisation des modèles SA (adaptés avec une régression linéaire à maximum de vraisemblance) donne lieu en général à un grand nombre de fausses alarmes. Pour réduire cet effet, un score de pénalité est appliqué à toutes les branches de mots clés afin de réduire leur valeur acoustique.

Dans [Peskin et al. 1996], l'objectif est la classification thématique sur le corpus de messages téléphonique Switchboard. La transcription est réalisée avec leur système de reconnaissance (avec un taux d'erreur de 40 %) qui utilise un ensemble de modèles de langages thématiques. Les ML attribuent des scores pour trouver le thème. Leur méthode est à base d'une petite liste de mots clés ou d'expressions sélectionnées avec le test χ^2 qui mesure le taux de distinction d'un mot par rapport à un message.

Chapitre III

Classification thématique

1 Introduction

L'identification thématique est la classification automatique de segments de textes ou messages de parole dans un des ensembles connus des thèmes possibles. C'est donc un problème de classification où la tâche est l'assignement d'un label thématique correct à un message connu comme appartenant à un nombre fixé des thèmes possibles.

La classification en thèmes suscite de nombreux travaux en recherche documentaire. Plusieurs méthodes ont été développées pour la classification de documents écrits. Récemment, des méthodes statistiques fondées sur les modèles unigrammes ont été proposées, notamment dans [Li et Yamamishi 1997]. Par ailleurs, la classification thématique fait l'objet de nombreux travaux en RAP (voir [Chen et al. 1998], [Carlson 1996], [Imai et al. 1997], [Peskin et al. 1996], [Seymore et al. 1998], [Wright et al. 1996]). Pour la transcription automatique de documents de parole ([Wright et al. 1996], [Carlson 1996], [Peskin et al. 1996]), les mélanges de modèles statistiques sont souvent employés ([Imai et al. 1997] ; [Li et Yamamishi 1997]).

Ce chapitre introduit un nouveau modèle stochastique du langage pour la classification thématique de textes écrits, et du discours. La classification thématique est un processus de décision appliqué à un document dont le résultat est l'assignation d'une étiquette thématique choisie parmi un ensemble donné d'étiquettes possibles. En pratique, un document peut appartenir à plusieurs thèmes. Par conséquent, il peut arriver qu'un document soit constitué d'une séquence de segments que l'on pourrait classer différemment. C'est pourquoi, nous nous intéresserons ici à des segments de textes plus courts, dont l'unité est le paragraphe.

Pour construire une représentation statistique des thèmes, il est nécessaire de disposer d'un ensemble de données préalablement triées thématiquement. Une interrogation qui se pose alors, et pour laquelle on ne trouve que peu de réponses dans la littérature, est la définition d'un *thème*.

Les dictionnaires classiques (Hachette, Petit Larousse et Petit Robert) proposent les définitions énumérées ci-après :

- *Hachette* :
  1. **thème :** 1. Sujet, matière, proposition que l'on entreprend de traiter dans un ouvrage, un discours. *Quel est le thème de cet essai ?* Syn. sujet 2. LING Syn. topique.
  2. **sujet :** Ce qui donne lieu à la réflexion, à la discussion ; ce qui constitue le thème principal d'une oeuvre intellectuelle, artistique. *Sujet de conversation. Le sujet d'une thèse, d'un tableau.*
  3. **topique :** LING (Anglicisme) Personne ou chose dont on dit quelquechose (par oppos. à commentaire, ce qui est dit de cette personne ou chose). Syn. thème
- *Petit Larousse* :
  1. **thème :** Sujet, idée sur lesquels portent une réflexion, un discours, une œuvre autour desquels s'organise une action. *Le thème d'un débat.*
  2. **sujet :** Matière sur laquelle on parle, on écrit, on compose. *Le sujet d'une conversation, d'un film.*
- *Petit Robert* :
  1. **thème :** Sujet, idée, proposition qu'on développe. Syn. sujet
  2. **sujet :** Ce qui, dans une oeuvre littéraire, constitue le contenu d'une pensée sur lequel s'est exercé le talent créateur de l'auteur. *Un sujet de roman, un bon sujet, un sujet en or.*
  3. **topique :** qui se rapporte exactement au sujet dont on parle. *Argument topique.*

Ces définitions manquent de précision pour notre objectif et diffèrent sensiblement selon l'éditeur. De plus, il n'y a pas de notion de niveau d'abstraction, en effet, il y a une différence importante de niveau entre des thèmes comme *"le sport"*, ou *"le football"*, ou encore *"la coupe du monde de football"*, ou bien *"Finale de la coupe du monde de football 1998"*, etc... La section 2 présente l'ensemble des étapes que nous avons utilisées pour la segmentation du corpus en unités thématiques. Les intitulés et codes des thèmes qui sont présentés dans la table III.1.

1	Etranger	5	Affaire, Economie
2	Histoire	6	Culture, Arts, Livres, Media
3	Sciences	7	Politique
4	Sports		

TAB. III.1 – Les 7 thèmes retenus

Dans ce chapitre, nous décrivons deux modèles complémentaires (sections 3 et 4). Le premier est composé d'unigrammes thématiques basés sur tous les mots du vocabulaire. Le second modèle repose sur une mémoire cache et des distributions statiques des mots clés thématiques. Au fur et à mesure du document, on effectue la comparaison entre la fenêtre temporelle représentée par la mémoire cache et les mots clés, ceci, pour chacun des thèmes. Dans le cas où les modèles assignent des labels thématiques différents pour le même texte, nous appliquons une règle de décision (section 5). Les différents résultats obtenus sont présentés en section 6.

2 Constitution des corpora thématiques

La création de modèles de langages thématiques nécessite un ensemble de corpora thématiques qui donneront une représentation statistique des thèmes. Ces données seront utilisées pour les apprentissages et pour les différents tests. Nous avons extrait le corpus à partir de deux cédéroms du journal *Le Monde* (1987 à 1991), ce qui représente un total de 80 millions de mots avec un vocabulaire de plus de 500 000 mots. Les thèmes de ces articles ne sont pas connus, mais, nous avons utilisé les secteurs rédactionnels du journal. Plusieurs étapes sont ensuite nécessaires afin d'obtenir un corpus thématiquement cohérent.

Nettoyage

En ce qui concerne l'accentuation, le LIA a décidé d'utiliser le format BDLEX[12] pour une grande partie de ses corpora. Le but d'un apprentissage est de modéliser les phénomènes linguistiques. Il est donc nécessaire de disposer d'un corpus débarrassé des passages ayant des valeurs linguistiques nulles ou spécifiques comme : les en-têtes, les lignes de titre, ainsi que les informations concernant les images qui sont intégrées dans les articles. Par ailleurs, un traitement préalable à cette segmentation consiste à repérer les débuts de paragraphes, que l'on signale par l'insertion d'une balise particulière : <p>.

Extraction

Nous avons utilisé 2 cédéroms du journal *Le Monde*, qui couvrent l'actualité des années 1987 à 1991. Ces données sont classées en trente secteurs de rédaction. Après avoir supprimé les unités trop petites pour constituer un thème, et avoir regroupé celles qui paraissent très proches comme par exemple "Sciences" et "Le monde des sciences", quinze catégories émergent : *Affaires*, *Arts*, *Biographie*, *Culture*, *Divers*, *Economie*, *Election*, *Etranger*, *Histoire*, *Livre*, *Media*, *Nécrologie*, *Politique*, *Sciences* et *Sport/Loisirs*. Nous avons supprimé les catégories *Biographie* et *Nécrologie* car tous les articles qui les composaient se retrouvaient dans d'autres catégories. Par la suite, nous avons supprimé le thème *Divers* car il provient de la fusion de plusieurs secteurs de rédaction différents et ne représente pas une unité thématique spécifique. A l'issu de ces choix, il est donc resté douze thèmes.

Segmentation en mots

Le texte est segmenté en phrases et en entités lexicales. Cette segmentation doit être cohérente avec les unités utilisées dans les traitements suivants, et reste un problème du fait de l'absence de normalisation. Nous nous retrouvons confrontés à des problèmes tels que :

1. le découpage des mots composés, par exemple *anti-naturel*, *Paris-Marseille*,
2. le regroupement de locutions linguistiques comme *parce que*, *tout de suite*,
3. la segmentation de nombres tels *31,5 %*.
4. la segmentation en phrases.

Le LIA utilise son propre outil de segmention. Le résultat est un document qui contient un mot par ligne, et indique respectivement le début et la fin de phrase par les marqueurs <s> et </s>.

[12]Représentation des accents sur 2 caractères. Par exemple é devient e1, è devient e2, ê devient e3, ï devient i4 et ç devient c5.

2.1 Analyse de la cohérence des regroupements thématiques

Pour vérifier la cohérence des données, il faut que les thèmes soient différents les uns des autres, et que chaque thème constitue réellement une "unité". Pour ce faire, nous avons développé deux approches ; la première porte sur l'étude du vocabulaire de chacun des thèmes tandis que la seconde intègre la fréquence des mots.

Etude du vocabulaire

Cette approche permet de comparer les vocabulaires des différents thèmes avec deux types d'évaluations. On calcule d'abord le taux d'inclusion de T_i dans T_j, par rapport à T_i, qui s'exprime par la formule :

$$\frac{\mid T_i \cap T_j \mid}{T_i}$$

qui est le nombre de mots de l'intersection d'un thème T_i avec un thème T_j, par rapport au vocabulaire de T_i. De même, on calcule le rapport du nombre de mots de l'intersection de T_i avec T_j, par rapport à leur union avec :

$$\frac{\mid T_i \cap T_j \mid}{\mid T_i \cup T_j \mid}$$

Les résultats que nous obtenons ([Bigi 1997]) sur nos corpora nous laissent supposer que les vocabulaires des thèmes suivants sont proches :

(a) *Affaires* avec *Economie*,

(b) *Culture* avec *Arts*, *Livre* et *Media*,

(c) *Politique* avec *Elections*.

Afin de valider la méthode, nous avons segmenté séquentiellement le corpus d'un thème en quatre sous-thèmes et nous avons observé leurs comportements entre-eux, et par rapport à d'autres thèmes. Nous avons pu observer que le vocabulaire d'un thème est généralement divisé en 3 groupes :

- un vocabulaire de base commun à tous les thèmes B (déterminants, adverbes...),
- un vocabulaire de base pour le thème T (mots clés du thème),
- un vocabulaire spécifique aux paragraphes S (mots qui n'apparaissent qu'exceptionnellement dans le thème).

Ainsi, plus le thème est grand, plus S augmente, tandis que B et T atteignent un "seuil". Ce qui nous incite à utiliser notre méthode sur l'ensemble du corpus, avec des sous-catégories de tailles identiques, afin que la variation de S ne soit pas observée. Ceci nous amène alors aux conclusions suivantes :

1. l'intersection de *Affaires* et *Economie* prend les mêmes valeurs que celle de *Affaires* avec *Affaires* : 55 % du vocabulaire de chacune des sous-catégories de *Affaires* est dans chacune des sous-catégories de *Economie*, et 55 % du vocabulaire des sous-catégories de *Affaires* sont dans les autres sous-catégories de *Affaires*.
2. La même constatation est faite avec les thèmes *Culture*, *Arts Livre* et *Media*, tandis que tous les autres ne laissent supposer aucune fusion.

Ces thèmes ont un vocabulaire proche et donc un regroupement peut être souhaitable. Ceci confirme les hypothèses (a) et (b) précédentes.

Etude de la fréquence des mots

Nous avons utilisé la formule de Bayes, avec T_i un thème et V_{T_i} son vocabulaire :

$$P(T_1|V_{T_2}) = \frac{P(V_{T_2}|T_1).P(T_1)}{P(V_{T_2})} \tag{III.1}$$

et :

$$P(V_{T_2}|T_1) = \prod_{i=1}^{|V|} \{w_i\}_{T_1} * n(w_i, T_2)$$

$$P(V_{T_2}) = \sum_{T_i}^{|T|} P(V_{T_2}|T_i) * P(T_i)$$

Nous avons fait varier $P(T_1)$ aux valeurs :

- $P(T_1) = \frac{1}{|T|}$,
- $P(T_1) = \frac{|V_{T_1}|}{\sum_{i=1}^{|T|} |V_{T_i}|}$.

si :

- $|T|$ est le nombre de thèmes,
- $|V_T|$ est la taille du vocabulaire d'un thème T,
- $n(w_i, T_2)$ est le nombre d'occurrences du mot w_i dans T_2
- $\{w_i\}_{T_1}$ est la i-ème composante du vecteur de probabilités des mots w_i dans T_1, telle que :

$$\{w_i\}_{T_1} = \frac{n(w_i, T_1)}{\sum_{j=1}^{|V_{T_1}|} n(w_j, T_1)}$$

sachant que si le mot n'a pas d'occurrence, on utilise une valeur seuil fixe $\{w_i\}_{T_1} = \varepsilon$.

Les résultats obtenus avec cette méthode nous montrent que l'écart entre les thèmes est une fonction croissante de la taille du corpus traité. Ils confirment les regroupements (a) et (b).

2.2 Le corpus final

Afin de constituer un corpus de test indépendant du corpus d'apprentissage, nous avons retiré un vocabulaire de 20 000 mots à 11 des 12 thèmes décrits précédemment. En effet, *Election* est supprimé car il est de taille insuffisante. Après avoir effectué les fusions (a) et (b), nous obtenons le corpus d'apprentissage décrit dans la table III.2. On constate que les différents thèmes sont inégalement représentés.

Thèmes	Etranger	Histoire	Science	Sport	Affaires Economie	Culture Arts Livre Media	Politique
Taille Vocabulaire	172 943	38 447	66 842	16 730	167 098	274 602	132 962
Nb Mots Corpus	24 293 083	609 703	2 025 741	191 966	20 796 734	25 253 457	13 442 692

TAB. III.2 – Taille du corpus d'apprentissage

L'ensemble des étapes présentées dans cette section ont été décrites plus en détails dans le mémoire de DEA [Bigi 1997].

3 Unigrammes thématiques

Un modèle de langage unigramme associe une probabilité à chaque mot du vocabulaire. Dans le cas d'un unigramme thématique, les probabilités des mots dépendent du thème. Celles-ci sont obtenues à partir des corpora thématiques décrits dans la section précédente.

Lors de la classification thématique d'un segment de texte, on calcule $P_1(T_j \mid W_1^{i-1})$, la probabilité d'un thème connaissant l'histoire, où W_1^{i-1} est la séquence $\{w_1, ...w_{i-1}\}$ des premiers $(i-1)$ mots d'un document. Cette probabilité accumule toutes les informations de l'historique de la manière suivante :

$$P_1(T_j \mid W_1^{i-1}) = \frac{P(T_j)P(W_1^{i-1} \mid T_j)}{\sum_{k=1}^{J} P(T_k)P(W_1^{i-1} \mid T_k)} \tag{III.2}$$

où J est le nombre total de thèmes et $P(T_j)$ est la probabilité du thème T_j calculée comme suit :

$$P(T_j) = \frac{\sum_w n(w, T_j)}{\sum_k^J \sum_w n(w, T_k)}$$

avec $n(w, T_j)$ qui est le nombre d'occurrences du mot w dans le thème T_j et :

$$P(W_1^{i-1} \mid T_j) = \prod_{t=1}^{i-1} P_1(w_t \mid T_j)$$

Un des reproches que l'on peut faire à l'utilisation des modèles unigrammes est qu'ils ne prennent pas en compte l'apport sémantique des mots. Le manque de données d'apprentissage est également un problème fréquemment rencontré lors de leur estimation. Une solution possible est l'introduction de méthodes pour ré-estimer les probabilités des mots, et notamment pour donner une estimation aux mots inconnus.

Les méthodes de backing-off ou d'interpolation (décrites dans le chapitre IV), peuvent être utilisées à cette fin. Selon Pereira, Tishby et Lee [Peirera et al. 1993], la similarité de 2 mots w_a et w_b, suivant un vocabulaire V, peut être mesurée en utilisant la distance de Kullback-Leibler ([Cover et Thomas 1991]) telle que :

$$D(w_a, w_b) = \sum_{w \in V} P(w \mid w_a) \log \frac{P(w \mid w_a)}{P(w \mid w_b)}$$

Les modèles standard de backing-off sont utilisés quand la probabilité bigramme ne peut pas être estimée à partir des données. L'ensemble $S(w_a)$ des mots similaires à w_a est l'ensemble des mots w_b pour lesquels $D(w_a, w_b)$ est plus petite qu'un seuil approprié. L'idée de base est que les mots qui ont des distributions n-grammes similaires pourraient être sémantiquement proches et peuvent être regroupés dans une classe. Ces concepts vont être étendus dans la section suivante en y intégrant une mémoire cache.

4 Modèle cache

L'utilisation d'une mémoire cache a été présentée la première fois par Kuhn et De Mori ([Kuhn et De Mori 1990]). Une mémoire cache correspond à une fenêtre de largeur M qui précède le mot courant et qui se déplace sur le texte.

L'utilisation d'informations thématiques nécessite un ensemble de mots caractéristiques du thème qu'ils représentent et que l'on appellera *mots clés*. Il est possible de trier les L mots du vocabulaire selon leur fréquence dans le corpus d'apprentissage et de sélectionner automatiquement les K plus fréquents pour chaque thème T_j. Cependant, certains mots ne sont pas de bons candidats, comme par exemple les adjectifs, les nombres ou encore les déterminants. Dans une phase de pré-traitement, bon nombre d'entre eux ont été identifiés en sélectionnant les mots les plus probables de V et en les triant manuellement. Ils ont été placés dans une liste de manière à ce qu'ils ne fassent pas partie des mots clés. Cette stop liste contient 1388 entrées.

4.1 Distance de Kullback-Liebler symétrique

On appelle $K(T_j)$ l'ensemble des mots clés du thème T_j. Leur distribution statistique est obtenue à partir du corpus d'apprentissage. Elle est continuellement comparée au contenu d'une mémoire cache de taille m, dont G mots sont différents. Ce processus est décrit en figure III.1.

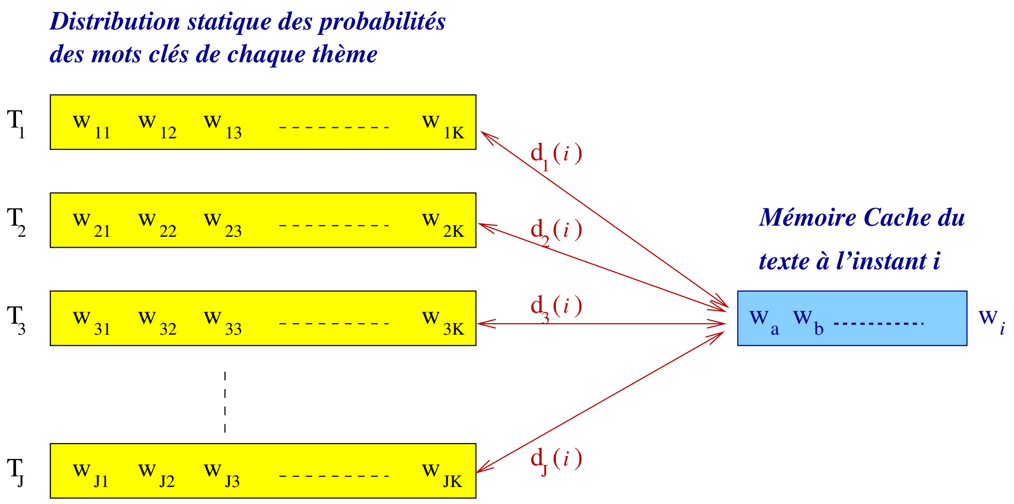


FIG. III.1 – Fonctionnement du modèle cache pour la classification thématique

Ainsi, à chaque instant i, pour chaque thème T_j, on calcule une distance $d_j(i)$, entre deux distributions de probabilités associées aux données suivantes :

- les mots du texte qui sont entrés dans la mémoire cache,
- les mots clés du thème.

Pour éviter que les mots qui ne sont pas représentatifs des thèmes interviennent dans la décision, il est important de faire en sorte que tous les mots qui ne sont pas dans le cache aient une probabilité proportionnelle à une probabilité d'un unigramme général. Ce même seuil sera alors attribué aux mots qui ne sont pas des mots clés du thème. De cette manière, tous les mots qui ne sont pas mots clés apporteront la même contribution, quel que soit le thème, et quel que soit le texte. Leur contribution à la distance doit être nulle. C'est pourquoi, seuls les mots-clés, indépendemment de leur thème d'origine, entrent dans le cache.

L'ensemble de ces définitions donnent la contribution pour le cache C_i d'un mot w à l'instant i telle que :

$$P_{cache}(i,w) = \begin{cases} \beta\frac{n_i(w)}{m_i} & si \quad w \in C_i \\ \alpha P_g(w) & si \quad w \notin C_i \end{cases} \tag{III.3}$$

où :
- $n_i(w)$ est le nombre d'occurrences du mot w dans le cache C_i au temps i ;
- m_i est le nombre total d'items dans le cache au même moment ;
- α et β sont des coefficients de normalisation ;
- $P_g(w)$ est la probabilité seuil associée à tous les mots qui ne sont pas dans le cache.

On peut noter que la somme des occurrences des $n_i(w)$ pour tous les mots du cache est égale à m_i.

La probabilité associée à la distribution des mots clés, pour un mot w du thème T_j est :

$$P_j(w) = \begin{cases} \gamma_j P_1(w \mid T_j) & si\ w \in K(T_j) \\ \alpha P_g(w) & si\ w \notin K(T_j) \end{cases} \tag{III.4}$$

où :
- $P_1(w \mid T_j)$ est la probabilité unigramme du mot w dans le thème T_j ;
- γ_j est un coefficient de normalisation dépendant du thème T_j ;
- $P_g(w)$ est la même probabilité seuil que dans l'équation III.3, associée à tous les mots qui ne sont pas mots clés du thème T_j.

L'évaluation de la distance commence quand le cache atteint un seuil minimal de mots. Lors de l'insertion dans le cache du n-ème mot du texte, la distance de KL symétrique est calculée comme suit :

$$d_j(n) = \sum_{i=1}^{K(T_j)+m} (P_{cache}(n,w_i) - P_j(w_i))log\left(\frac{P_{cache}(n,w_i)}{P_j(w_i)}\right) \tag{III.5}$$

où $d_j(n)$ est la distance relative au thème j après avoir lu les n premiers mots. Ce calcul implique quatre situations :

1. $(w \in C_i) \wedge (w \in K(T_j))$
2. $(w \in C_i) \wedge (w \notin K(T_j))$
3. $(w \notin C_i) \wedge (w \in K(T_j))$
4. $(w \notin C_i) \wedge (w \notin K(T_j))$

Comme il a été prévu, tous les mots qui ne sont mots clés dans aucun des thèmes ont la probabilité seuil à la fois dans la distribution dépendante du thème et dans le cache (cas 4). Ainsi, leur contribution à la distance de KL est nulle. Ces mots n'ont donc pas besoin d'entrer dans le cache.

La dernière étape à réaliser est la normalisation de la distance ainsi obtenue soit après l'entrée d'un mot clé dans la mémoire cache ou à la fin d'une phrase s, pour chaque thème :

$$d_j^*(s) = \frac{d_j(s)}{d_j(0)} \tag{III.6}$$

4.2 Contraintes sur les coefficients

Les coefficients α, β, γ et la valeur de $P_g(w)$ permettent de normaliser les distributions de probabilités afin qu'elles somment à 1. Ceci implique un ensemble de contraintes décrites ci-après.

Contrainte sur γ_j

On définit les deux notations suivantes :

1. $P_{KT_j} = \sum_{k=1}^{K(T_j)} P_1(w_k \mid T_j) = \frac{\sum_{k=1}^{K(T_j)} n_j(w_k)}{\sum_{l=1}^{L(T_j)} n_j(w_l)}$ est la somme des probabilités des mots clés du thème T_j, avec :
   - $L(T_j)$ est le vocabulaire de T_j,
   - $n_j(w)$ est le nombre d'occurrences de w dans T_j.
2. $P_{\overline{K(T_j)}} = \sum_{w \notin KT_j} P_g(w)$ est la somme des probabilités seuil des mots qui ne sont pas mots clés du thème T_j.

On a donc la propriété $P_{KT_j} + P_{\overline{K(T_j)}} = 1$ qui doit être respectée. La contrainte sur γ_j issue de la formule III.4 s'exprime alors comme :

$$\gamma_j P_{KT_j} + \alpha P_{\overline{KT_j}} = 1$$

d'où l'on déduit γ_j tel que :

$$\gamma_j = \frac{1 - \alpha P_{\overline{KT_j}}}{P_{KT_j}} = \frac{1 - \alpha(1 - P_{KT_j})}{P_{KT_j}} = \frac{1 - \alpha + \alpha P_{KT_j}}{P_{KT_j}}$$

Contrainte sur $P_g(w)$

$P_g(w)$ est une probabilité seuil que l'on donne aux mots qui ne sont pas dans le cache, ou qui ne sont pas dans la distribution de mots clés lors du calcul de la distance. Elle doit donc être fixée afin qu'elle soit plus petite que la probabilité qu'un mot appartienne au cache, donc $P_g(w) < \frac{1}{m}$. Il faut également que $P_g(w)$ soit plus petite que la plus petite des probabilités d'un mot clé dans un thème, soit :

$$P_g(w) < \min_j \left[\min_{i \in K(T_j)} (P_1(w \mid T_j)) \right]$$

Cette valeur est obtenue expérimentalement, la table III.3 montre les résultats qui ont été obtenus pour la plus petite probabilité d'un mot clé dans chaque thème. L'observation des nombres repostés ci-dessous a permis de fixer $P_g = 10^{-5}$.

Etranger	0,0000133
Histoire	0,0000148
Sciences	0,0000142
Sports	0,0000156
Economie	0,000013
Culture	0,0000133
Politique	0,0000126

TAB. III.3 – Plus petites probabilités observées pour un mot clé dans chaque thème

Contrainte sur α et β

On a tout d'abord des contraintes dues à la formule III.3 :

$$\beta \sum_{w \in C_i} \frac{n_i(w)}{m_i} + \alpha \sum_{w \notin C_i} P_g(w) = 1$$

Mais la somme des probabilités des mots du cache est égale à 1 donc :

$$\beta + \alpha \sum_{w \notin C_i} P_g(w) = 1$$

$$\beta = 1 - \alpha \sum_{w \notin C_i} P_g(w)$$

On a également des contraintes sur les unigrammes. La plus petite occurrence d'un mot dans le cache est 1, il faut que ce mot clé ait une probabilité plus grande que les non mots clés :

$$\frac{\beta}{m_i} > \alpha P_g(w_i)$$

En principe, $\sum_{w \notin C_i} P_g(w)$ est dépendante du temps car le contenu de la mémoire cache varie. En pratique, ce terme est toujours très proche de 1 car la mémoire cache contient un nombre de mots bien inférieur à L. Il est possible de l'approximer par sa valeur maximale et de considérer :

$$\beta = 1 - \alpha \sum_{MAX} P_g$$

4.3 Optimisation du calcul de la distance de KL

Lors du calcul de la distance (formule III.5) on effectue $K(T_j) + m$ itérations qui correspondent au calcul de $d_j(n)$ avec tous les mots clés du thème T_j et ceux contenus dans la mémoire cache qui ne sont pas dans le thème. Or, seul le cache évolue dans le temps, les mots clés étant déterminés à l'avance ainsi que leur distribution statistique. Ainsi, à chaque calcul d'une nouvelle distance, la majorité des itérations sont identiques à celles du calcul de la distance au temps 0 (cache vide) : seuls les mot clés entrés dans le cache donnent des itérations différentes.

Ces redondances dans les évaluations peuvent être évitées si l'on considère que la distance de KL à un instant i est égale la distance de KL à cache vide à laquelle on ajoute la participation des mots qui sont entrés dans le cache (et on enlève la contribution de ces mêmes mots à la distance à cache vide). Pour cela, il faut commencer par évaluer la distance à cache vide en début de programme. Ensuite, chaque calcul de la distance est effectué de la façon suivante :

$$d_j(i) = d_j(0) + \sum_{k \in cache, k \in T_j} (A(w_k) - B(w_k)) + \sum_{k \in cache, k \notin T_j} C(w_k) \tag{III.7}$$

où :

1. $A(w)$ est le cas où le mot qui est dans le cache est dans T_j (i.e. $(w \in C_i) \wedge (w \in K(T_j))$) :

$$A(w) = \left(\beta \frac{n_i(w)}{m_i} + \gamma_j P_1(w \mid T_j)\right) + log\left(\frac{\beta \frac{n_i(w)}{m_i}}{\gamma_j P_1(w \mid T_j)}\right)$$

2. $B(w)$ est le cas où le mot est dans T_j mais n'est pas dans le cache (c'est la contribution d'un mot pour le calcul de la distance à cache vide) (i.e. $(w \notin C_i) \wedge (w \in K(T_j))$) :
$$B(w) = (\alpha P_g(w) + \gamma_j P_1(w \mid T_j)) + log\left(\frac{\alpha P_g(w)}{\gamma_j P_1(w \mid T_j)}\right)$$
3. $C(w)$ est le cas où le mot qui est dans le cache n'est pas dans T_j (i.e. $(w \in C_i) \wedge (w \notin K(T_j))$) :
$$C(w) = \left(\beta\frac{n_i(w)}{m_i} + \alpha P_g(w)\right) + log\left(\frac{\beta\frac{n_i(w)}{m_i}}{\alpha P_g(w)}\right)$$

L'application de cette formule permet donc de n'effectuer qu'une seule fois les $K(T_j)$ itérations. Par la suite, à chaque calcul de la distance, seulement m itérations seront réalisées. Ceci permet un gain de temps de calcul assez important, qui rend le processus de classification thématique en fonctionnement "temps réel".

4.4 Des distances normalisées aux probabilités du modèle cache

Plusieurs fonctions peuvent être appliquées aux distances normalisées afin obtenir la répartition de probabilités thématiques du modèle cache. La contrainte à respecter est d'utiliser une fonction qui fera en sorte que la meilleure distance (la plus petite) donne le meilleur score (le plus grand). Pour ce faire, si $X = d^*_j(s)$, issu de la formule III.5, on peut utiliser par exemple $\exp(X)$ ou $\frac{\alpha}{X^2}$. Par la suite, une simple normalisation permet d'obtenir la distribution de probabilités thématiques.

5 Règle de décision

Dans la mesure où le modèle unigramme et le modèle cache peuvent assigner des étiquettes thématiques différentes, l'idée est d'associer un indice de confiance à chacun des modèles en fonction du contexte qui entoure la décision, tel que le nombre de mots clés du cache, ou encore la probabilité du meilleur thème dans les modèles unigrammes. L'indice de confiance est noté μ.

Dans un premier temps, il faut associer un seuil de confiance au cache en dessous duquel on ne lui accorde pas la prise de décision. Ainsi, quand un nouveau texte est traité, aucune connaissance *a priori* n'est donnée sur son thème, et ce n'est qu'après quelques mots v_1 que l'on peut commencer à faire des hypothèses sur son thème. La décision thématique est basée sur l'écart des distances du meilleur thème et du deuxième meilleur, sous réserve que le cache contienne un nombre suffisant de mots. Ceci se traduit par :

$$\mu_1(s) = \begin{cases} 0 & \textit{si le cache contient moins de } v_1 \textit{ mots} \\ d^*_{2b}(s) - d^*_m(s) & \textit{sinon} \end{cases} \tag{III.8}$$

où :

$d^*_m(s) = \min_j d^*_j(s)$, la distance obtenue par le meilleur thème,

$d^*_{2b}(s) = \min_{j,j\neq m} d^*_j(s)$, la distance obtenue par le deuxième meilleur.

La décision prise à l'instant s dépend de l'écart des probabilités des deux meilleurs thèmes du modèle cache mais aussi du nombre de mots dans le cache et de la valeur de la meilleure probabilité unigramme. Une fonction binaire est ainsi donnée telle que :

$$\mu_2(s) = \begin{cases} 1 & si\ \{\mu_1(s) > \upsilon_2\} \cap \{[\max_j P_1(W_1^s \mid T_j)]\} > \upsilon_3\} \\ 0 & sinon \end{cases}$$

Un degré de plausibilité est enfin attribué au modèle cache :

$$\rho_2(s) = \mu_1(s) \wedge \mu_2(s)$$

et celui du modèle unigramme est $\rho_1(s) = 1 - \rho_2(s)$, auquel cas, la décision est basée sur le thème dont $P_1(W_1^s \mid T_j)$ est maximale. Cette règle de décision, issue de la logique floue, est également présentée dans l'article [Bigi et al. 2000a].

6 Résultats

Un corpus de test de 1 021 paragraphes a été extrait parmi les articles du journal "Le Monde" des années 1987 à 1991. Ce corpus n'a pas été utilisé dans la phase d'apprentissage. Il représente 7 576 phrases, 221 335 entités lexicales. Les valeurs que nous avons déterminées (par expérimentations) comme pertinentes pour les différents paramètres du modèle cache sont :

1. les coefficients de normalisation : $\alpha = 0,1$ et $\beta = 0,9$,
2. le nombre de mots clés par thème est $K(T_j) = 4000$,
3. la taille du cache G est de 100 mots clés,
4. le cache est utilisé après $\upsilon_1 = 5$ mots clés.

La table III.4 donne les résultats de la classification thématique du corpus de test, en comparant le label assigné automatiquement à un label de référence qui a été attribué manuellement. Les probabilités sont évaluées mot après mot et les résultats présents sont basés sur les labels thématiques assignés à la fin de chaque paragraphe. Dans la première ligne de la table :

- U représente le thème retenu par l'unigramme seulement,
- C représente le thème retenu par le modèle cache,
- S représente le thème assigné manuellement au segment de texte, qui constitue ici la solution de référence,
- C out représente le cas où le cache contient moins de 5 mots clés,
- le symbole = représente l'accord,
- le symbole $\neq$ représente le désaccord.

On remarque que 586 paragraphes (57,4 %) sont correctement étiquetés à la fois par le modèle cache et le modèle unigramme tandis que $108 + 111 = 219$ (21,45 %) des paragraphes sont étiquetés correctement par un seul des deux modèles. Ce sont sur ces derniers que la règle de décision s'appliquera (les résultats sont en section III.6.4).

6.1 Variation du nombre de mots clés par thèmes

Les expériences réalisées (III.5) montrent que dans notre cas le nombre de mots clés peut varier entre 4000 et 5000.

Thème	$U \neq S$ $C \neq S$	$U = S$ $C \neq S$	$U \neq S$ $C = S$	$U = S$ $C = S$	$U \neq S$ C out	$U = S$ C out	Total
Etranger	16	23	9	81	4	6	139
Histoire	0	1	0	6	0	0	7
Sciences	19	8	19	43	2	0	91
Sports	22	14	53	20	2	3	114
Economie	6	7	6	124	3	13	159
Culture	33	35	14	262	8	29	381
Politique	36	20	10	50	7	7	130
Total	132	108	111	*586*	26	58	1021

TAB. III.4 – Résultats de classification thématique, en nombre de paragraphes (chaque ligne correspond au secteur de rédaction d'origine du paragraphe)

Nombre de mots clés	$U = C = S$	C out	Score Unigramme	Score Cache	Règle de décision
3000	572	70	752 (73,65%)	751 (73,55%)	824 (80,71 %)
3800	576	62	752 (73,65%)	744 (72,87%)	819 (80,22 %)
4000	586	58	752 (73,65%)	755 (73,95%)	824 (80,71 %)
4200	594	57	752 (73,65%)	762 (74,63%)	823 (80,61 %)
5000	601	52	752 (73,65%)	770 (75,42%)	822 (80,51 %)
6000	601	49	752 (73,65%)	758 (74,24%)	817 (80,02 %)

TAB. III.5 – Résultats de classification thématique en fonction du nombre de mots clés sélectionnés par thème (stop liste de 1388 mots)

6.2 Variation du nombre de mots de la stop liste

La table III.6 montre les résultats en faisant varier le nombre de mots de la stop liste. La deuxième ligne correspond à l'anti-dictionnaire fourni par le Laboratoire Parole et Langage (LPL) d'Aix-en-Provence[13], et la 3ème ligne est la stop liste que nous avons créée (LIA). Les résultats indiquent que la stop liste est utile dans le modèle, mais elle n'affecte que très peu les résultats.

Nbre de mots stopliste	U=C=S	Cout	Score Unigramme	Score Cache	Règle de décision
0	583	58	752 (73,65%)	748 (73,26%)	821 (80,41%)
684 (LPL)	591	58	752 (73,65%)	753 (73,75%)	822 (80,51%)
1388 (LIA)	586	58	752 (73,65%)	755 (73,95%)	824 (80,71%)

TAB. III.6 – Résultats de classification thématique en fonction du nombre de mots de la stop liste (en utilisant une sélection de 4000 mots clés par thème)

[13]Disponible à l'adresse : http ://www.up.univ-mrs.fr/~veronis/data/antidico.txt

6.3 Utilisation des lemmes pour la classification

Etiquetage morpho-syntaxique

L'étiquetage morpho-syntaxique de surface, proposé par T. Spriet et M. El Bèze dans [Spriet et El-Bèze 1998], détermine la classe (souvent syntaxique) de chacun des mots qui composent le message (nom masculin singulier, pronom personnel...). Cette classe permet alors de lever l'ambiguïté pour de nombreuses applications (homophonie, ré-accentuation, sémantique). Les outils linguistiques destinés au traitement des langues naturelles doivent également disposer d'un module robuste de gestion de l'inconnu. Les mots hors-vocabulaire peuvent être classés selon plusieurs catégories (noms propres, faute de frappe, fautes d'orthographe, néologisme, mots composés, entités extra-linguistiques, etc...).

Le tagger du LIA choisit une classe pour chacun des mots de la phrase, parmi une liste de 103 classes morphologiques possibles (citées en Annexe B). Cet étiquetage est réalisé par une méthode statistique mixée avec une approche à base de règles. L'étiquetage intervient aussi dans des domaines connexes aux textes écrits tels que la reconnaissance de la parole, ou la synthèse. Différents modules ont été intégrés à cet outil :

- deux modules qui cherchent les classes probables, selon des méthodes statistiques en fonction des finales des mots, un arbre de décision et un modèle 3-lettres,
- une approche triviale pour la recherche des noms propres, elle se contente de déclarer toutes les classes nom propre possibles (nom de famille, nom de ville féminin, nom de pays masculin...) et laisser le tagger choisir en fonction du contexte de la phrase,
- une petite étude à base de règles pour le cas des mots vus dans des contextes particuliers (M. Xinconnu, rue Quelquechose, etc...).

Lemmatisation

Le lemmatiseur utilise la classe morphologique affectée à un mot pour, si nécessaire, lever l'ambiguïté sur son lemme de rattachement. La table III.7 est un exemple de phrase telle qu'elle est traitée. L'accentuation est au format BDLEX, la première colonne est le mot, la deuxième représente sa classe obtenue avec le tagger cité précédemment et la troisième son lemme.

La forme lemmatisée d'un texte peut être utilisée pour sa classification thématique de la même manière qu'on le fait avec les mots. Nous avons créé un ensemble d'unilemmes thématiques, c'est à dire que l'on évalue la probabilité d'apparition de la forme lemmatisée des mots dans chaque thème. On utilise donc $P_1(l_w|T_j)$ au lieu de $P_1(w|T_j)$ où w_l représente le lemme du mot w. Pour le modèle à base de mémoire cache, nous avons sélectionné un ensemble de lemmes clés qui remplaceront les mots clés dans le modèle. Ce nombre a été fixé arbitrairement à 3000. Le fonctionnement du modèle reste le même : les lemmes clés entrent dans la mémoire cache et on calcule la distance entre ce contenu de la mémoire cache et l'ensemble de lemmes clés retenus pour chaque thème. La table III.8 donne les résultats de la classification des textes lemmatisés sur le corpus de test. On constate que les résultats sont très proches de ceux obtenus avec les mots.

<s>	ZTRM	<s>
La	DETFS	le
cite1	NFS	cite1
attend	V3S	attendre
,	YPFAI	,
comme	COSUB	comme
aux	PREPAUX	a2+le
premiers	AMP	premier
jours	NMP	jour
,	YPFAI	,
que	COSUB	que
les	DETMP	le
soleils	NMP	soleil
couchants	MOTINC	couchants
incendient	V3P	incendier
ses	DETFP	son
fac5ades	NFP	fac5ade
de	PREPADE	de
briques	NFP	brique
.	YPFOR	.
</s>	ZTRM	</s>

TAB. III.7 – Phrase extraite du corpus d'apprentissage, taggée et lemmatisée.

Thème	$U \neq$S $C \neq$S $U \neq$C	$U = S$ $C \neq S$ $U \neq C$	$U \neq S$ $C = S$ $U \neq C$	$U = S$ $C = S$ $U = C$	$U \neq S$ C out	$U = S$ C out	Total
Etranger	15	16	11	79	5	13	139
Histoire	0	2	0	5	0	0	7
Sciences	23	8	12	41	4	3	91
Sports	19	16	45	26	6	2	114
Economie	5	9	4	117	3	21	159
Culture	36	31	15	262	7	30	381
Politique	40	15	10	51	7	7	130
Total	136	102	104	*595*	27	57	1021

TAB. III.8 – Résultats de classification thématique des lemmes, en nombre de segments (classés par secteur de rédaction d'origine du texte)

6.4 Synthèse des résultats

La table III.9 résume les résultats de la classification thématique du jeu de test. La première ligne se réfère au cas où le cache et l'unigramme sont en accord avec le label manuel. Dans la deuxième ligne, on ajoute le cas où l'unigramme est en accord avec le label et le cache ne contient pas assez de données pour être pris en compte. La troisième ligne intègre à la deuxième, la décision prise par l'unigramme seulement. La quatrième ligne intègre à la deuxième, la décision prise par le cache seulement. La dernière ligne correspond à la règle de décision appliquée au cache et à l'unigramme.

	Mots		Lemmes		Combinés	
	N	%	N	%	N	%
$U = C = S$	586	57,39	561	54,95	595	58,28
$+C$ out	644	63,08	650	63,66	652	63,86
Deux premiers + Unigramme	752	73,65	754	73,85	754	73,85
Deux premiers + Cache	755	73,95	741	72,58	755	73,95
Stratégie de combinaison	824	**80,71**	819	80,22	828	81,1

TAB. III.9 – Synthèse des résultats de classification thématique de textes écrits

La tâche de la classification thématique est très délicate à traiter. Elle repose en effet sur la **décision humaine** et son caractère subjectif fait qu'elle est souvent remise en cause par d'autres personnes. Le modèle cache est meilleur avec l'utilisation des 4000 mots clés qu'avec 3000 lemmes clés. Par contre, les unilemmes apportent une légère amélioration par rapport aux unigrammes. La solution appelée "combinée" développe une approche hybride où les deux modèles sont le modèle cache qui utilise les mots clés, et des unilemmes thématiques. Mais les différences entrent ces résultats ne sont pas suffisamment significatives pour qu'elles puissent être exploitées.

Les résultats montrent une augmentation des performances grâce à l'utilisation du cache. Avec la contribution du cache, une précision supérieure à 80 % est obtenue. Ces expériences montrent donc les avantages de l'utilisation conjointe d'un modèle cache et d'un modèle unigramme. Leur complémentarité permet d'aboutir à une amélioration substantielle de la classification thématique.

Pour donner une idée du résultat proposé par le modèle cache, la figure III.2 montre l'exemple d'un paragraphe court extrait du corpus de test du thème sport. Le texte traité est le suivant (les mots clés sont indiqués en caractères gras) :

" Depuis ***dix-sept*** *ans que je dispute le* ***Safari Rally*** *, je n' ai jamais connu* ***ça*** *. Après les* ***pluies*** *qui se sont abattues sur le pays depuis un mois , l' état des* ***routes*** *et des* ***pistes*** *est inimaginable , surtout dans les monts Taita " ,* ***affirmait*** *le* ***Suédois*** *Bjorn Waldegaard (* ***Toyota Celica*** *GT 4) , trois fois* ***vainqueur*** *de l'* ***épreuve*** *en 1977, 1984 et 1986 .*

Parmi l'ensemble des mots clés de ce paragraphe, seul le mot *pluies* n'est pas un mot clé du thème sport. Le modèle cache commence à donner ces valeurs après le mot *pluies*, qui est le 5ème mot clé, ce qui correspond à la 22ème entité lexicale. Ceci est illustré par la figure III.3

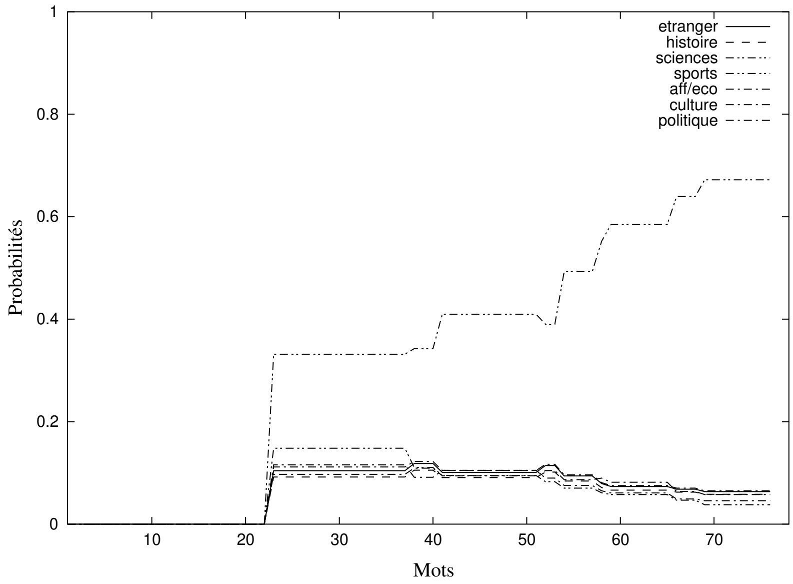


FIG. III.2 – Exemple de classification thématique donnée par le modèle cache

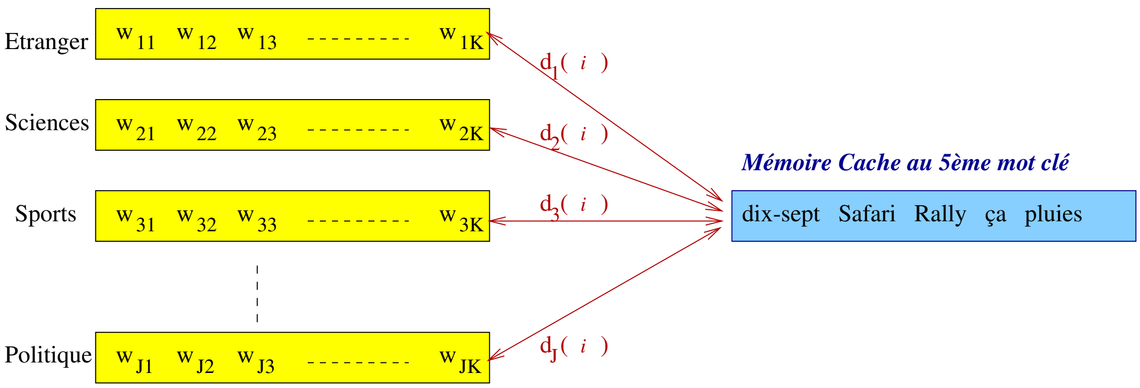


FIG. III.3 – Représentation du modèle cache au 5ème mot clé du paragraphe de la figure III.2

6.5 Plusieurs étiquettes thématiques

Lors de l'analyse thématique d'un paragraphe, il arrive que plusieurs thèmes soient abordés (politique étrangère, sport et économie...). Il est alors délicat de choisir quel thème attribuer au segment. Les traitements proposés précédemment, ne tiennent pas compte de ce phénomène pourtant fréquent. C'est pourquoi, lors de l'étiquetage manuel du corpus de test, deux labels thématiques ont été assignés aux paragraphes. Les résultats précédemment cités ne se référaient qu'au premier de ces 2 labels. Cependant, lorsqu'il n'apparaît aucune ambiguïté sur le thème, le même label est assigné deux fois.

Sur les 1021 paragraphes du corpus de test, 678 ont deux fois le même label thématique, ce qui représente les 2/3 des données. Les résultats obtenus lors de la classification de ces segments sont présentés en table III.10. On observe de très bons taux de classification pour les deux modèles (79,5 % pour l'unigramme seul, et 83 % pour le modèle cache seul). En appliquant la règle de décision, on obtient plus de 89 % de classification correctes.

On en conclut que lorsqu'il n'y a pas d'ambiguïté sur le thème d'un texte, notre méthode de classification est encore plus performante.

	N	%
$U = C = S$	441	64,95
$+C$ out	483	71,13
Deux premiers + Unigramme	539	*79,5*
Deux premiers + Cache	564	*83,06*
Règle de décision	607	**89,53**

TAB. III.10 – Classification thématique des paragraphes pour lesquels un seul thème représente la solution (678 paragraphes)

Les résultats de classification des paragraphes pour lesquels 2 labels thématiques différents représentent la solution sont en table III.11. On compare le thème choisi par chacun des modèles avec les deux solutions possibles. Celles-ci sont notées par $S1$ et $S2$, où $S1$ est la solution comparée lors des évaluations précédentes et $S2$ le deuxième label thématique qui est affecté.

		N	%
	$U = C = S1$	145	42,27
	$+C$ out	161	46,94
(a)	Deux premiers + Unigramme	212	61,81
(b)	Deux premiers + Cache	191	55,69
	Règle de décision	220	64,14
(c)	(a) + $U = S2$	285	83,09
(d)	(b) + $C = S2$	278	81,05

TAB. III.11 – Résultats de la classification thématique lorsque deux labels différents sont des solutions possibles (343 paragraphes)

La ligne (c) présente le cas où l'unigramme a assigné l'un des 2 labels solution et la dernière ligne (d) présente le cas où le modèle cache a assigné l'un des 2 labels solution. On observe également de très bons taux de classification pour les deux modèles (83 % pour l'unigramme seul, et 81 % pour le modèle cache seul). La figure III.4 donne un exemple de paragraphe dont les deux labels sont différents : Culture ($S1$) et Science ($S2$). Le texte de ce paragraphe est le suivant :

DONNER au grand public une information ***scientifique sérieuse*** *qui soit à la portée du plus grand nombre : tel est le but des* ***éditions*** *Le Rocher, qui ont lancé* ***récemment*** *une nouvelle* ***collection*** *: "****Sciences*** *et* ***découvertes****". D'un* ***format*** *de* ***poche*** *(150X178mm), chaque* ***ouvrage****,* ***rédigé*** *par un* ***spécialiste*** *du* ***sujet traité****, compte 128 ou 160* ***pages*** *et 10 à 30* ***illustrations*** *noir et blanc au* ***trait****. Pas de* ***photos*** *ou d'****images*** *en* ***couleurs*** *(sauf sur la* ***couverture****). Cette* ***austérité*** *a son avantage : le prix de chaque livre est de 35 ou de 39 francs. A ce jour, une* ***douzaine*** *de* ***titres*** *sont* ***disponibles****. Ils* ***sortent****, en effet, au rythme de deux toutes les six semaines. Leur* ***variété*** *est très grande : de la* ***Vie*** *des* ***étoiles*** *aux Oasis du fond des* ***mers*** *et des Dinosaures à la Moisissure, du Cerveau hormonal aux Volcans et magmas...*

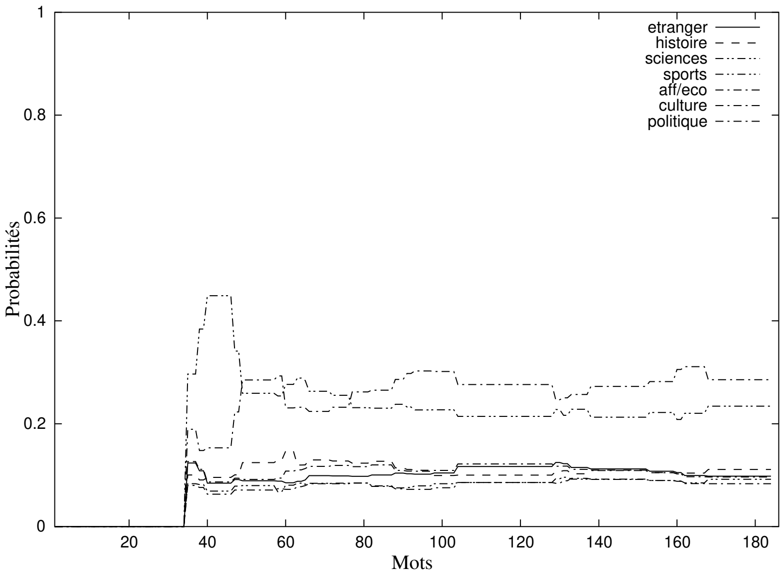


FIG. III.4 – Exemple de classification thématique donnée par le modèle cache, les deux labels solution sont Culture et Sciences

6.6 Classification thématique de documents de parole

La classification de documents de parole trouve naturellement ses applications en recherche documentaire, mais aussi, en reconnaissance automatique de la parole, comme nous le verrons dans le chapitre suivant. Dans cette section, nous décrivons l'utilisation de l'outil de classification, en l'utilisant sur des textes dictés. Pour évaluer le taux de classification thématique de textes dictés, on compare le label assigné par notre système au texte original à celui assigné au texte dicté (sortie du système de reconnaissance). Le corpus de test composé de 97 paragraphes (18000 mots) a été dicté au système *ViaVoice 98* d'IBM, un système dépendant du locuteur et à grand vocabulaire. La répartition en thèmes de ce corpus est présentée dans la table III.12. Quatre locuteurs (1 femme, 3 hommes) ont participé à sa constitution. Nous avons observé sur ce corpus un taux d'erreurs d'environ 35 %.

Etranger	17	Economie	15
Histoire	4	Culture	17
Sciences	15	Politique	14
Sports	15		

TAB. III.12 – Composition du corpus dicté : nombre de paragraphes dictés par thèmes

Nous obtenons 73 paragraphes correctement labélisés contre 77 sur les textes de référence correspondants. Comme on pouvait le penser, les mots clés sont apparemment bien reconnus. Ces résultats sont détaillés dans la table III.13. Ils confirment que nos travaux de classification thématique sur l'écrit peuvent être réutilisés dans le cadre de retranscription de documents de parole.

Thème	$E \neq S$ $D \neq S$ $E \neq D$	$E \neq S$ $D \neq S$ $E = D$	$E = S$ $D \neq S$	$E \neq S$ $D = S$	$E = S$ $D = S$	$E \neq S$ D out
Etranger	1	2	2	1	10	1
Histoire	1		1		2	
Sciences	1		3	3	8	
Sports		1	2		12	
Economie		1	1	1	11	1
Culture	1		1	2	13	
Politique		2	2	1	9	
Total	4	6	*12*	*8*	**65**	2

TAB. III.13 – Résultats de la classification thématique sur des textes dictés. S représente le label de référence, E celui donné par le modèle cache sur le texte écrit, et D celui donné par le modèle cache sur le texte dicté.

Pour donner une idée du comportement du modèle cache sur un texte dicté, la figure III.5 montre l'exemple d'un paragraphe court extrait du corpus de test du thème sport. Le texte transcrit par le système de dictée est le suivant :

*" Depuis 17 ans . Que je dispute le **assassin riz** et **rallyes** , je n' ai jamais connu sa . Après les **pluies** qui se sont abattues sur le pays depuis un mois , l' état des **routes** et des **pistes** inimaginables , surtout dans l' aimant **habitat** " , **affirmé** le **suédois** décent **à l' égard de** (**Toyota Celica** GT4) , trois fois **vainqueur** tollé preuve en 1977, 1984 et que 1986 .*

Parmi l'ensemble des mots clés de ce paragraphe, les mots suivants ne sont pas des mots clés de sport : *assassin, riz, pluies, habitat, affirmé, à l'égard de* n'est pas un mot clé du thème sport. Le modèle cache commence à donner ces valeurs après le mot *routes*, qui est le 5ème mot clé, ce qui correspond à la 42ème entité lexicale.

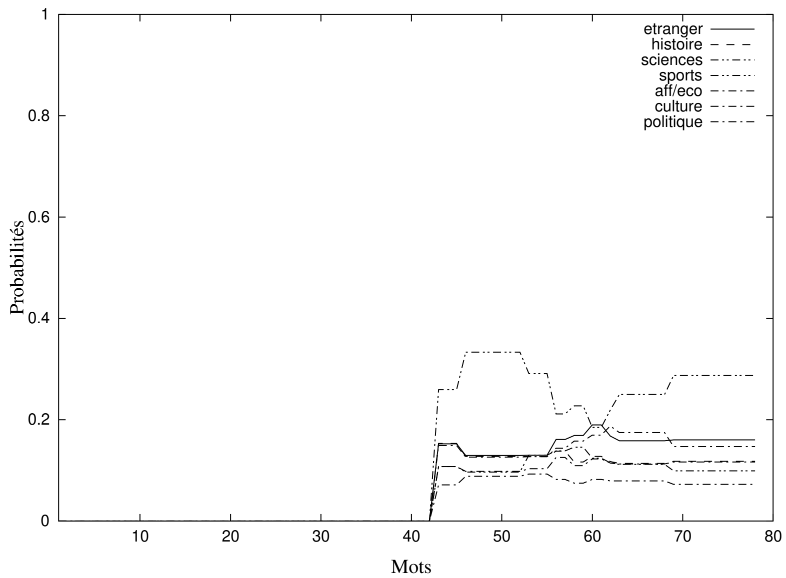


FIG. III.5 – Exemple de classification thématique par le modèle cache sur un texte dicté

7 Perspectives

Dans ce chapitre, nous avons développé de nouveaux concepts en modélisation statistique du langage. Nous avons introduit une méthode basée sur une mémoire cache, avec une sélection de mots clés thématiques. Le modèle repose sur la comparaison entre la distribution statistique des mots contenus dans la mémoire cache d'un texte à un instant donné et les distributions statiques des mots clés des thématiques. Cette évaluation est réalisée avec la divergence de Kullback-Liebler symétrique, et évolue dans le temps avec la prise en compte de nouveaux mots dans le cache. Nous avons vu que cette méthode est performante sur des textes écrits, et qu'il est possible de reconnaître des thèmes au fur et à mesure qu'un document est dicté.

Les résultats sur la classification en thèmes de documents écrits montrent d'une part la bonne performance de la mémoire cache et d'autre part son rôle complémentaire à celui des unigrammes thématiques pour la procédure de décision. Les résultats de la classification de documents dictés, valident également la méthode. Ils montrent la capacité du modèle cache à s'adapter à de nouvelles tâches. Les avantages offerts par le modèle cache ont également donné la possibilité d'affecter plusieurs labels thématiques aux segments de textes. Appliqués à la Reconnaissance Automatique de la Parole, les résultats entrevus dans ce chapitre permettent d'envisager l'utilisation d'une combinaison linéaire de modèles de langage thématiques lors de la phase de reconnaissance. Nous pouvons également envisager de comparer l'histoire de la dictée d'un document avec des distributions thématiques et d'en utiliser les résultats pour l'adaptation dynamique du langage (chapitre IV). Ces méthodes de comparaison dynamique peuvent aussi être adaptées à la segmentation thématique de textes (chapitre V), et à l'expansion de requêtes (chapitre VI) dans le cadre d'une application de recherche documentaire.

Chapitre IV

Adaptation des modèles de langage

1 Introduction

Les méthodes d'adaptation des modèles de langage ont pour but de capter les différents degrés de changement du langage. Pour l'adaptation thématique, les modèles peuvent être appris sur des corpora qui représentent des domaines spécifiques (rapport médicaux, correspondance commerciale...). Ces domaines n'utilisent qu'une partie du langage, un sous-langage, d'un point de vue du lexique, de la syntaxe, de la sémantique et de la structuration du discours ; ceci implique des conséquences sur les modèles de langage. L'impact peut être favorable sur les performances des systèmes de reconnaissance s'ils sont dédiés au domaine pour lequel leurs modèles de langage ont été appris. Mais les performances risquent d'être fortement dégradées si, par exemple, on voulait dicter une lettre commerciale à un système prévu pour des rapports médicaux ! Les techniques d'adaptation tentent de capter les dépendances longue distance et/ou d'ajuster les statistiques des modèles n-grammes dynamiquement en repérant les préférences de l'utilisateur et les changements thématiques. D'un point de vue de l'application et de l'évaluation, plusieurs types d'adaptation sont possibles. D'une façon générale, l'adaptation peut être vue comme un problème d'estimation paramétrique du modèle.

Ce chapitre présente en premier lieu la théorie d'estimation d'un modèle optimal, avec plusieurs estimations possibles, dont l'estimation bayésienne qui s'applique bien à l'adaptation des modèles à un nouveau domaine. Ce chapitre aborde ensuite le problème de l'estimation des données manquantes. En effet, dans la mesure où les corpora d'apprentissage ne constituent pas une énumération de toutes les formes de la langue, les modèles de langage ignorent certaines formes du langage qui pourraient constituer des suites de mots acceptables. L'estimation de ces données manquantes est donc une partie importante dans la création des modèles. Par la suite, sans chercher à être exhaustif, nous décrivons quelques modèles de langage en rapport avec l'adaptation, à savoir, les modèles dynamiques, les modèles distants, les modèles trigger et le modèle cache, puis les modèles thématiques. La dernière section présente les travaux que nous avons réalisés en matière d'adaptation. L'objectif est de déterminer automatiquement le thème d'un texte au cours de sa dictée et d'introduire la participation du modèle de langage correspondant dans le système de reconnaissance. Pour valider ce procédé, nous montrons la puissance prédictive des modèles thématiques lorsque le thème a été déterminé automatiquement.

2 Estimation des probabilités des modèles

L'apprentissage des modèles de langage dépend directement du type de modèle utilisé. Pour un modèle purement statistique, comme celui des n-grammes, il s'agit d'un problème d'estimation des probabilités. Considérons un ensemble de *textes d'apprentissage* $W = w_1 \ldots w_I$ constitués de mots appartenant à un vocabulaire fini V. A partir de l'hypothèse que l'historique h_i d'un mot w_i est limitée aux $n-1$ mots, l'ensemble des textes réservés à cet effet peut être appris, sans perte d'informations, de la façon suivante :

$$S = h_i w_i, h_{i+1} w_{i+1}, \ldots, h_I w_I$$

Pour un historique h donné, un corpus S_h peut être extrait de S en prenant la sous-séquence de tous les n-grammes dans S qui commencent par h, on a alors :

$$S_h = h w_{h_1}, \ldots, h w_{h_m}$$

où $\{h_1, h_2, \ldots, h_m\} \subseteq \{n, n+1, \ldots, I\}$ peut être vue comme une suite de m mots extraits indépendamment selon la probabilité $P(w \mid h)$. En d'autres termes, le corpus S_h est constitué de m variables aléatoires de *distributions identiques et indépendantes* (IID). Par la suite, nous allons nous limiter à un contexte donné h, lors de la présentation de l'ensemble des méthodes utilisées pour estimer les probabilités conditionnelles $\{P(v \mid h)\}, v \in V$, d'un modèle M. Il suffit de procéder de la même manière pour estimer les probabilités conditionnelles $\{P(v \mid h')\}$, utilisant d'autres contextes tels que $h' \neq h$. Pour des raisons de simplicité, le contexte h sera omis des notations. L'objectif est l'estimation de la distribution de probabilités discrètes $P(v), v \in V$, donnée par le modèle M à partir d'un corpus d'apprentissage $S = w_1, \ldots, w_m$ de variables aléatoires de distributions identiques et indépendantes. Il faut noter que v fait référence à un mot du vocabulaire, alors que w_i représente le i-ème mot de S. La distribution $P(v)$ appartient généralement à une famille paramétrique :

$$\{P(v; \theta), \theta \in \Theta\}$$

où θ est un vecteur de paramètres inconnus qui spécifie la distribution, et Θ est l'espace de paramètres. Dans ce qui suit, nous nous limiterons à l'étude de deux distributions : l'une appelée *discrète* et l'autre *discrète symétrique.*

2.1 Distribution Discrète

L'espace de paramètres de la distribution discrète est défini comme suit :

$$\Theta = \{\theta = [\theta_v]_{v \in V} : \theta_v \geq 0 \quad \text{et} \quad \forall v \in V, \sum_{v \in V} \theta_v = 1\}$$

qui assigne directement un paramètre θ_w à chaque mot v du vocabulaire V. En fait,

$$P(v; \theta) = \theta_v$$

La vraisemblance du corpus $P(S; \theta)$ est définie ainsi :

$$P(S; \theta) = \frac{m!}{\prod_{v \in V} c(v)!} \prod_{v \in V} \theta_v^{c(v)} \quad \text{où} \quad c(v) = \sum_{i=1}^{m} \delta(w_i = v) \tag{IV.1}$$

avec $\delta(e) = 1$ si e est vraie, 0 sinon. Le terme m est le nombre de symboles dans S et le vecteur $[c(v)]_{v \in V}$ représente les statistiques suffisantes du corpus S.

2.2 Distribution Discrète Symétrique

Si on suppose, par besoin de symétrie, que les mots ayant la même fréquence dans S doivent avoir la même probabilité, on définit alors une distribution paramétrique légèrement différente, telle que :

$$P(v;\theta) = \frac{\theta_r}{n_r} \quad \text{où} \quad r = c(v)$$

où θ_r représente la probabilité totale de tous les mots de $v \in V$ présents r fois dans S, alors que n_r est le nombre de mots présents r fois dans ce même S. Un tel modèle introduit moins de paramètres que les partitions de w.

La vraisemblance du corpus S avec la distribution symétrique est par conséquent :

$$P(S;\theta) = \frac{m!}{\prod_{v\in V} c(v)!} \prod_{v\in V} \left(\frac{\theta_{c(v)}}{n_{c(v)}}\right)^{c(v)} = \frac{m!}{\prod_{v\in V} c(v)!} \prod_{r\geq 0} \left(\frac{\theta_r}{n_r}\right)^{rn_r} \tag{IV.2}$$

Pour estimer la valeur de probabilité $P(v;\theta)$, il suffit de définir la meilleure valeur pour son vecteur de paramètres Θ. Le problème de l'estimation des points est un problème classique en statistiques paramétriques et peut être abordé de différentes façons. Plusieurs critères différents d'estimation sont possibles. Nous présentons dans ce qui suit l'estimation par maximum de vraisemblance et l'estimation bayésienne.

2.3 Estimation par maximum de vraisemblance

Les probabilités du modèle sont estimées par un critère qui considère le paramètre θ comme une quantité inconnue à déterminer. Dans cette approche, la meilleure estimation est définie comme étant celle qui maximise la probabilité (ou la vraisemblance) du corpus d'apprentissage S. L'estimation de probabilité maximale (MP) de θ, suivant le principe du maximum de vraisemblance est définie comme suit :

$$\begin{aligned}
\theta^{MP} &= \arg\max_{\theta\in\Theta} P(S;\theta) \\
&= \arg\max_{\theta\in\Theta} \prod_{i=1}^{m} P(w_i;\theta)
\end{aligned} \tag{IV.3}$$

L'estimation MP d'une distribution discrète peut être facilement évaluée en maximisant le logarithme de la vraisemblance plutôt que la vraisemblance elle-même. En prenant ceci en compte pour le calcul de Θ, et en supprimant un facteur constant, on obtient la fonction lagrangienne suivante :

$$L(\theta;\lambda) = \sum_{v\in V} c(v) \log\theta_v + \lambda\left(1 - \sum_{v\in V}\theta_v\right)$$

En dérivant ceci par rapport à θ_v, $v \in V$ et en ramenant la fonction à zéro, on obtient :

$$\frac{\partial L}{\partial\theta_v} = \frac{c(v)}{\theta_v} - \lambda = 0 \qquad \forall v \in V$$

Après un réarrangement des termes, cette formule devient :

$$\sum_{v\in V} c(v) = \left(\sum_{v\in V}\theta_v\right)\lambda \;=\; \lambda$$

En substituant λ, on obtient l'estimation suivante :

$$\theta_v^{MP} = \frac{c(v)}{\sum_{v \in V} c(v)} = \frac{c(v)}{m} \tag{IV.4}$$

Dans le cas d'une distribution symétrique, si on maximise le logarithme de la vraisemblance par rapport à θ, on a alors :

$$\theta_r^{MP} = \frac{n_r r}{m} \tag{IV.5}$$

ce qui produit les mêmes probabilités que la distribution discrète.

2.4 Estimation bayésienne

L'estimation bayésienne s'applique bien à l'adaptation à un nouveau domaine d'un modèle appris dans un domaine. En fait, le vecteur de paramètres θ est considéré comme une variable aléatoire pour laquelle la distribution *a priori* est supposée connue. Dans ce cas, le problème est de trouver une estimation de θ, étant donné le corpus d'apprentissage S et la distribution *a priori* $P(\theta)$ des paramètres de θ. En appliquant la règle de Bayes, la distribution *a posteriori* de θ est définie comme suit :

$$P(\theta \mid S) = \frac{P(S \mid \theta)\ P(\theta)}{P(S)} \tag{IV.6}$$

La distribution *a posteriori* de θ combine l'existant *a priori* avec l'ensemble empirique du corpus d'apprentissage. Les deux expressions différentes de la probabilité, $P(S \mid \theta)$ et $P(S; \theta)$, expriment la même distribution. Cette différence de notation permet de mettre en évidence le fait que Θ est un paramètre dans $P(S; \theta)$, tandis que Θ est considéré comme une variable aléatoire dans $P(S \mid \theta)$. L'estimation de θ peut être dérivée de la distribution *a posteriori* par différentes méthodes dont deux d'entre elles sont maintenant énumérées.

1. Le critère du **Maximum A Posteriori (MAP)** cherche à faire en sorte que la valeur de θ maximise la probabilité *a posteriori*. En éliminant le facteur constant $P(S)$ dans IV.6, le critère MAP apparait comme une généralisation du MP :

   $$\theta^{MAP} = arg \max_{\theta \in \Theta} P(\theta \mid S) = arg \max_{\theta \in \Theta} P(S \mid \theta) P(\theta)$$

   Le plus simple, mais qui donne moins de connaissances *a priori* pour le vecteur θ est la distribution uniforme :

   $$P(\theta) = s(\theta) = \frac{1}{vol(\Theta)} \tag{IV.7}$$

   où $vol(\Theta)$ indique le volume de l'espace de paramètres Θ. Avec cette distribution, le MAP estimé est équivalent à celui de MP.

2. Le critère **Bayesien (B)** considère θ au regard de la distribution *a posteriori* :

   $$\begin{aligned} \theta^B &= E[\theta \mid S] \\ &= \int_\Theta \theta P(\theta \mid S) d\theta \\ &= \frac{\int_\Theta \theta P(S \mid \theta) P(\theta) d\theta}{\int_\Theta P(S \mid \theta) P(\theta) d\theta}. \end{aligned} \tag{IV.8}$$

   Une condition suffisante pour le calcul de la distribution *a posteriori* de IV.6 est que S ait des statistiques suffisantes, ce qui est garanti si la forme paramétrique $P(v; \theta)$ est de la *famille exponentielle*. En fait, l'existence de statistiques suffisantes implique l'existence de distributions $P(\theta)$ pour lesquelles la distribution *a posteriori* est de la même famille que l'antérieure.

3 Estimation des données manquantes

Dans un corpus d'apprentissage des modèles de langage, il est courant que toutes les suites de mots possibles ne soient pas rencontrées, alors que certaines d'entre elles peuvent constituer une suite de mots correcte, ou acceptable dans le langage. Les modèles tels que nous les avons présentés dans les sections précédentes attribuent une probabilité nulle à chaque séquence non observée dans la phase d'apprentissage, et donc refusent l'apparition de certaines suites possibles de mots. Le problème est le même pour les suites de mots peu probables qui seront éliminées au profit d'autres suites plus fréquentes. Une solution consiste à estimer la probabilité conditionnelle $\hat{P}(w \mid h)$ en combinant deux composantes : le modèle de lissage et le modèle de redistribution.

Le modèle de lissage est lié au problème de l'estimation des suites de mots dont la fréquence est nulle ([Witten et Bell 1991]). Dans ce modèle, la probabilité de tous les mots non observés dans un historique h est estimée en escomptant les fréquences des suites de mots observées. La fréquence relative d'un mot w étant donné son historique h est estimée suivant le principe du maximum de vraisemblance :

$$f(w \mid h) = \frac{c(hw)}{c(h)}$$

où $c(.)$ est le nombre d'occurrences de l'argument dans le corpus, et, quand $c(hw) = 0$, alors $f(w \mid h) = 0$. Le modèle de lissage produit ainsi une fréquence conditionnelle réduite $f^*(w \mid h)$ pour les mots tels que $c(hw) \neq 0$:

$$0 \leq f^*(w \mid h) \leq f(w \mid h), \qquad \forall\ hw \in\ V^n$$

où V est le vocabulaire et n le nombre de mots dans la suite de mots hw. On trouve dans la littérature plusieurs méthodes qui proposent de calculer le terme $f^*(w \mid h)$. Ces méthodes d'escompte sont discutées dans [De Mori 1998]. Nous citons dans les paragraphes suivant la méthode de Good-Turing, la méthode nommée repli (*backing-off*) et l'interpolation linéaire.

3.1 Méthode de Good-Turing

Dans cette méthode, le terme $f^*(w \mid h)$ est défini comme suit :

$$f^*(w \mid h) = \begin{cases} \frac{c(hw)^*}{c(h)} & si\ c(hw) > 0 \\ 0 & sinon \end{cases} \tag{IV.9}$$

et :

$$c(hw)^* = (c(hw) + 1).\frac{t_{c(hw)+1}}{t_{c(hw)}}$$

Le terme t_c indique le nombre de suite de mots différentes qui apparaissent c fois dans le corpus d'apprentissage. La distribution de probabilités des suites de mots non observées se définit à partir de la probabilité suivante :

$$\lambda(h) = 1 - \sum_{w \in V} f^*(w \mid h)$$

Dans le cas de la méthode de Good-Turing, cette équation devient :

$$\lambda(h) = 1 - \sum_{w:c(hw)>0} \frac{c^*(w \mid h)}{c(h)}$$

La probabilité $\lambda(h)$ est redistribuée sur tous les mots non rencontrés et dont l'historique est h. Cette redistribution des masses escomptées est réalisée proportionnellement à une distribution moins spécifique $P(w \mid h')$, où h' indique un historique réduit par rapport à h. Par exemple, dans le cas d'un modèle trigramme, c'est la distribution bigramme qui est utilisée.

3.2 Backing-off

Cette approche est proposée par Katz dans [Katz 1987]. Dans cette méthode, on estime la distribution $\hat{P}(w \mid h)$ telle que :

$$\hat{P}(w \mid h) = \begin{cases} f^*(w \mid h) & \text{si } f^*(w \mid h) > 0 \\ \alpha_h \lambda(h) P(w \mid h') & \text{sinon} \end{cases} \quad \text{(IV.10)}$$

où α est un facteur de normalisation (qui assure la condition $\sum_w \hat{P}(w \mid h) = 1$) défini comme suit :

$$\alpha_h = \left(\sum_{w:f^*(w|h)=0} P(w \mid h') \right)^{-1}$$

Dans le cas de l'estimation des modèles n-grammes, la distribution qui est utilisée quand la fréquence $f^*(w \mid h)$ est nulle est de type $(n-1)$-gramme.

3.3 Interpolation linéaire

Cette approche, proposée par F. Jelinek et R. Mercer ([Jelinek et Mercer 1980]), combine directement les distributions de probabilités, comme suit :

$$\hat{P}(w \mid h) = f^*(w \mid h) + \lambda(h) P(w \mid h')$$

Par exemple, si l'on veut obtenir l'interpolation linéaire d'un trigramme avec un bigramme, on obtient :

$$\hat{P}(w \mid h) = (1 - \lambda(h)) P_{trig}(w \mid h) + \lambda(h) P_{big}(w \mid h')$$

Les approches de backing-off et d'interpolation linéaire qui viennent d'être présentées utilisent récursivement les distributions d'ordre inférieur dans l'estimation du modèle global.

4 Modèles avec adaptation

4.1 Modèles dynamiques et modèles distants

Le besoin d'aller au-delà des modèles trigrammes est évoqué dans [Jelinek 1991]. Puisque ces modèles de langage utilisent un contexte très limité, ils ne peuvent pas détecter les variations de style ou de thème du texte. Les modèles à longues distances essaient de capter les dépendances qui dépassent les trigrammes. Différentes approches sont apparues, elles permettent soit l'utilisation des n-grammes pour $n > 3$, soit de considérer les variations de distributions de $P(w_i \mid h_i)$ en regardant les 100 derniers mots de h_i distants de w_i. Ces modèles sont souvent appelés dynamiques parce qu'ils modifient dynamiquement leurs statistiques en s'appuyant sur l'histoire.

Etant donné un ensemble de modèles, chacun représentant un domaine, il est possible de construire un nouveau modèle en combinant leurs distributions de probabilités et d'adapter dynamiquement les différents poids. Une possibilité pour les combiner consiste en une interpolation linéaire des différents paramètres comme suit (en s'assurant que $\sum_j \lambda_j(t) = 1$) :

$$P(w_t \mid h_t) = \sum_{j=1}^{J} \lambda_j(t) P(w_t \mid h_t)$$

$$\lambda_j(t) = \frac{1}{M} \sum_{m=0}^{M} \frac{\lambda_j(t-1) P_j(w_{t-m} \mid h_{t-m})}{\sum_{j=0}^{J} \lambda_i(t-1) P_i(w_{t-m} \mid h_{t-m})}$$

Les *permugrammes* ([Schukat-Talamazzini et al. 1994]) sont une généralisation des modèles n-grammes. Ils sont obtenus par interpolation linéaire d'un grand nombre de modèles conventionnels de type bigrammes, trigrammes et d'ordre supérieur. Ces derniers agissent sur des permutations différentes et spécifiques de l'historique h_t du mot à prédire w_t. Cette approche permet de tenir compte des dépendances entre les séquences de mots non adjacents qui peuvent être représentées sans avoir recours à des n-grammes d'ordre élevé.

Les *bigrammes d'expressions* étendent les bigrammes de mots vers des bigrammes d'expressions courtes. E. Giachin propose des algorithmes qui recherchent les paires de mots adéquates qui peuvent être associées pour former de nouvelles compositions de mots ([Giachin 1995]). Ceci est répété jusqu'à ce que la perplexité du modèle ne s'améliore plus. Les paires de mots adjacentes qui contiennent beaucoup d'informations mutuelles peuvent être réunies ([Kenne et al. 1995]). Dans [McCandless et Glass 1994], une première grammaire est automatiquement inférée à partir d'un corpus d'apprentissage. Un algorithme essaie ensuite de réduire cette grammaire en fusionnant les séquences de mots et les non-terminaux (i. e. les expressions). Finalement, la grammaire est utilisée pour générer un modèle de classes/expressions.

Les permugrammes et les modèles bigrammes d'expressions se sont montrés efficaces pour les petits vocabulaires des applications de dialogue homme-machine dans lesquelles la caractéristique principale des expressions est qu'elles contiennent un nombre important d'expressions courtes mais qui ont une très grande fréquence d'apparition. Dans ce cas, de bien meilleures prédictions peuvent être obtenues en utilisant de grands n-grammes. Cependant, aucune de ces méthodes n'a été testée avec l'explosion de taille occasionnée par l'utilisation de modèles d'ordre supérieur sur de grands vocabulaires.

Les modèles n-grammes *à mixture* ([Kneser et Steinbiss 1993]) essaient de détecter des contraintes longues distances dans une phrase ou un paragraphe en interpolant différents modèles n-grammes sur des thèmes spécifiques. L'objectif est de donner des poids plus importants aux thèmes en cours de reconnaissance. Dans [Iyer et al. 1994] une variation de cette approche est proposée. En particulier, l'interpolation des modèles est effectuée au niveau de la phrase, et les modèles dépendant du sujet sont estimés sur des échantillons de paragraphes obtenus à partir d'un algorithme de segmentation du texte.

Bahl et ses coauteurs ont proposé une extension des n-grammes, basée sur un arbre de classification ([Bahl et al. 1989]). Ces modèles basés sur des arbres de classification peuvent être considérés dans le cadre général de la classification d'arbres, cette théorie est décrite en détail dans [Breiman et al. 1984]. L'utilisation d'un arbre de classification binaire permet de définir un ensemble de classes d'équivalence, chacune assignée à une feuille de l'arbre.

Partant de la racine, une question binaire permet de choisir une branche à chaque noeud de l'arbre, jusqu'à atteindre une feuille. A chaque feuille de l'arbre est attachée une distribution de probabilités d'observer le mot suivant. Plus précisément, si $f_{1 \cdots l-1}$ représente la feuille de l'arbre à laquelle on accède après les $l-1$ questions, on donne :

$$P(W_l \mid W_{1 \cdots l-1}) \approx P(W_l \mid f_{1 \cdots l-1})$$

Cette technique permet de définir les questions à poser ainsi que la structure de l'arbre de classification selon un critère d'entropie minimale sur les distributions attachées aux feuilles.

Les modèles n-grammes *redimensionnables* utilisent les méthodes qui effacent les statistiques n-grammes particulières pour réduire les besoins en mémoire des modèles, sans pour autant trop dégrader les performances ([Murveit et al. 1994], [Brugnara et Federico 1996], [Seymore et Rosenfeld 1996], [Kneser 1996]). Ces techniques ont principalement été appliquées aux trigrammes mais peuvent être étendues à des n-grammes d'ordre supérieur avec succès, en diminuant les modèles 4-grammes et 5-grammes ([Kneser 1996]).

4.2 Modèle trigger et modèle cache

Ce type de modèle a pour objet d'étendre l'historique tout en évitant le problème du manque de données dans le corpus. L'objectif est de capter les informations contenues dans l'historique lointain du mot à prédire. Le modèle cache trouve son fondement sur le fait que les mots contenus dans l'historique ont plus de chance de réapparaître.

Les modèles *triggers ou de collocation* utilisent des dépendances longues distances entre les paires de mots, qui sont corrélées de façon significative à l'intérieur du corpus d'apprentissage. En pratique, on recherche dans une fenêtre de largeur donnée les paires de mots ayant une forte information mutuelle. Les triggers sont introduits sous la forme de contraintes pour les modèles ([Rosenfeld et Huang 1992]), soit en les interpolant avec un modèle n-gramme ([Ney et al. 1994]). Des triggers peuvent être intégrés à la composante linguistique pour améliorer l'adaptation des modèles [Sarukkai et Ballard 1997].

Les modèles avec une *mémoire cache* de [Kuhn et De Mori 1990] ont été utilisés avec des modèles triclasses pour remplacer les probabilités d'un mot dans une classe. Ainsi, la probabilité associée à un mot w_i du modèle triclasse :

$$P(w_i \mid g_{i-2}g_{i-1}) \approx \sum_{g_i \in G} P(w_i \mid g_i) P(w_i \mid g_{i-2}g_{i-1})$$

est remplacée par :

$$P_{cache}(w_i \mid g_i) = \lambda f_{train}(w_i \mid g_i) + (1-\lambda) f_{cache}(w_i \mid g_i)$$

où la première fréquence relative est calculée sur le corpus d'apprentissage, alors que la seconde est basée sur le contenu de la mémoire cache des N derniers mots ($N = 200$) dont la catégorie lexicale est égale à g_i. Les paramètres d'interpolation $0 < \lambda < 1$ peuvent être estimés selon différentes méthodes.

Il existe plusieurs méthodes pour combiner les modèles triggers et les modèles cache avec d'autres modèles plus classiques ([Ney et al. 1994], [Tillman et Ney 1996]). Dans le cas d'une interpolation linéaire d'un trigramme, d'un trigger et d'un cache, la probabilité associée à un mot w_i est la suivante :

$$P(w_i \mid h) = (1 - \lambda_1 - \lambda_2) P_{n-gram}(w_i \mid w_{i-2}w_{i-1}) + \lambda_1 P_{cache}(w_i \mid h) + \lambda_2 P_{trigger}(w_i \mid h)$$

4.3 Modèles thématiques

Les modèles thématiques proposent une représentation de sous-domaines du langage. Certains thèmes peuvent être très généraux comme la politique, le sport, la culture, les sciences... ou très précis, comme dans [Seymore et al. 1998] ou [McDonough et Gish 1994] avec des sujets comme la pollution de l'air, la musique, le crime, le contrôle des armes, le service public, l'éducation publique, la vie familiale... Certains de ces modèles ont déjà été présentés, dans ce mémoire car ils sont utilisés pour la classification ou la segmentation thématique mais leurs applications sont plurielles :
- classification automatique,
- regroupement de thèmes proches,
- re-découpage automatique d'un thème en sous-thèmes,
- suivi de thématique,
- adaptation dynamique des modèles thématiques en RAP,...

Les modèles de langage n-grammes sont fréquemment employés par les systèmes de reconnaissance de la parole pour contraindre et guider la recherche. Les valeurs typiquement affectées à N varient de 2 à 4, malgré leur performance reconnue, ces modèles manquent d'information sur les contextes à long terme. Dans [Mahajan et al. 1999], les auteurs montrent que la puissance prédictive des modèles de langage de n-grammes peut être améliorée en utilisant des informations du contexte à long terme. Ils emploient des techniques de recherche documentaire pour généraliser l'information disponible concernant le contexte avec des modèles de langage thématiques. Ils montrent l'efficacité de cette technique avec des expériences sur le corpus des textes du "Wall Street journal". La méthode proposée peut réduire la perplexité du modèle de langage (trigramme + stemming) de 37 %, ce qui indique la puissance prédictive du modèle de langage thématique. Le problème avec ce type de méthode est que le thème doit être connu. Dans [Khudanpur et Wu 1999] un modèle longue distance dynamique qui incorpore des contraintes sur les conditions thématiques est utilisé dans un système de reconnaissance. Ces concepts vont être étendus dans la section suivante.

5 Adaptation dynamique des modèles thématiques

5.1 Principe

L'objectif de notre travail concernant l'adaptation des modèles de langage est l'intégration d'un outil de classification thématique en vue d'améliorer les performances des systèmes de reconnaissance automatique de la parole. Cette adaptation est réalisée en fonction des thèmes identifiés dynamiquement lors de la reconnaissance. Il consiste à utiliser un modèle généraliste classique au début de la reconnaissance afin d'initialiser le processus de classification thématique, et de se servir de ces données afin de déterminer le thème du texte. Le modèle de langage peut alors être adapté, par exemple par interpolation linéaire entre le modèle classique et les modèles thématiques, en fonction de cette connaissance sur la nature du texte.

Pour la réalisation de cet objectif nous proposons d'utiliser l'outil de classification décrit dans le chapitre précédent. Il respecte les différentes propriétés qui suivent. Tout d'abord, il est impératif que le thème puisse être trouvé non pas lorsque l'on dispose du document en entier mais au cours de son traitement. Pour la même raison, il est préférable que peu de données soient utiles pour obtenir le thème, afin que les modèles thématiques puissent être intégrés rapidement. Le principe du modèle cache respecte ces deux contraintes puisque le cache est une fenêtre qui se déplace sur le texte et que seulement 5 mots clés sont utiles pour déterminer le thème. Par ailleurs, un autre facteur qui peut être pris en compte lors de l'intégration d'un nouvel outil dans un système de RAP concerne les ressources informatiques nécessaires à son fonctionnement, et le temps de calcul. Concernant ces deux points, là aussi le modèle cache est satisfaisant car il ne nécessite qu'un ensemble de mots clés par thème et de leur distribution statistique, ce qui est très peu au regard des données utilisées par les modèles n-grammes. De plus, le thème est identifié en temps réel (moins d'une seconde pour un paragraphe complet). Le principe général de cette approche, en intégrant notre système d'identification thématique est illustré par la figure IV.1.

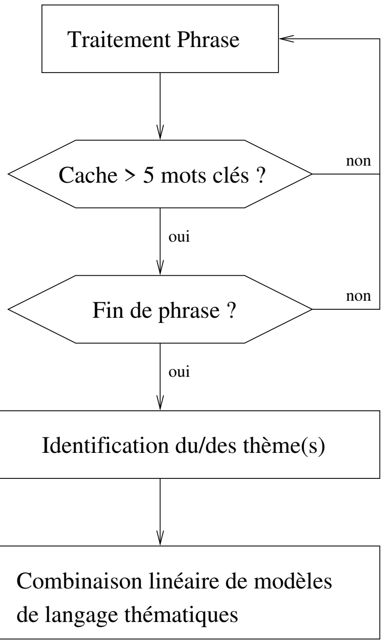


FIG. IV.1 – Détermination dynamique du thème des documents de parole

La validation de ce processus commence par l'évaluation de la capacité de prédiction des modèles de langage thématiques, par une mesure de perplexité. Idéalement, au vue de notre objectif, cette validation devrait avoir lieu sur des textes thématiques, résultat de la retranscription de textes de parole par un système de RAP. Ces textes doivent ensuite être classés en thèmes par le modèle cache tels qu'ils l'auraient été lors de leur dictée. Mais la constitution d'un tel corpus reste difficile est très fastidieuse. C'est pourquoi, à défaut de disposer des données adéquates, cette partie expérimentale sera effectuée sur des textes écrits. La deuxième étape de la validation consiste en l'intégration du système de classification thématique dans un système de RAP pour pouvoir aprécier le gain de reconnaissance. Ce travail n'a pas pu être conduit au cours de notre étude car le LIA ne disposait pas de système de reconnaissance de la parole au moment où ce travail a été conduit. Le travail que nous présentons dans les sections suivantes concernent donc l'évaluation des modèles de langage thématiques, en terme de perplexité.

5.2 Expérimentation

Dans cette section, nous proposons de valider l'idée qu'un modèle qui intègre des connaissances thématiques sera plus performant qu'un modèle généraliste, et ce, en déterminant automatiquement le thème du texte ([Bigi et al. 2000a]). Concernant le vocabulaire des modèles, si l'on veut que les modèles puissent être intégrés à un système de reconnaissance, il est préférable de le sélectionner de sorte à disposer des transcriptions phonémiques ainsi que des HMM de tous les mots qui le composent. Par ailleurs, la taille du vocabulaire ne doit pas être trop importante car les corpora d'apprentissage des modèles thématiques sont de taille limitée. C'est pourquoi, nous avons utilisé un sous-ensemble de 10 000 mots du vocabulaire du système MAUD, outils de traitement de la parole continue du LORIA ([Fohr et al. 1997], [Zitouni 2000]).

La création de l'ensemble des modèles et les calculs de perplexité ont été effectués avec le toolkit v2 du CMU ([Rosenfeld 1995]). Les tailles des corpora d'apprentissage dont nous disposons pour la création des modèles sont décrites dans le tableau IV.1. Chaque ligne du tableau indique le nombre d'entités lexicales disponibles dans le corpus pour chacun des 7 thèmes, et la dernière ligne concerne l'ensemble des données utilisées pour apprendre le modèle général. Comme le nombre de documents disponibles pour chaque thème est très différent, les données du modèle général ont été extraits aléatoirement des thèmes *Etranger, Economie, Culture et Politique.*

Thème	*Apprentissage*
Etranger	15 178 982
Histoire	609 703
Science	2 025 741
Sport	191 966
Economie	10 394 924
Culture	15 786 307
Politique	10 084 571
Général	32 341 193

TAB. IV.1 – Taille de l'ensemble des corpora utilisés pour entraîner les modèles.

On constate que certains thèmes ne disposent pas d'un corpus suffisant, c'est pourquoi, nous avons limité notre étude aux modèles bigrammes. Cependant, les thèmes *Sport, Histoire et Sciences* ne disposent pas de suffisamment de données pour apprendre un bigramme robuste. Nous avons alors choisi de ne pas utiliser les modèles thématiques directement mais de les interpoler avec un modèle mieux appris : le modèle général.

Nous utilisons la combinaison linéaire de chaque bigramme thématique avec le bigramme général, telle que :

$$P(w_i \mid w_j) = \lambda P_g(w_i \mid w_j) + (1 - \lambda) P_t(w_i \mid w_j)$$

où $P_g(w_i \mid w_j)$ et $P_t(w_i \mid w_j)$ sont les probabilités du mot w_i étant donné l'observation de w_j respectivement dans le modèle général et dans le modèle du thème t. Le calcul de la perplexité du modèle général et celui de chaque modèle combiné est alors réalisé sur le corpus de test séparé en thèmes par le modèle cache, comme décrit dans le chapitre III.

λ	Etranger	Histoire	Sciences	Sport	Economie	Culture	Politique
0,0	139,72	292,11	177,95	232	137,40	162,65	145,65
0,1	137,38	282,15	174,63	227,16	136,26	162,46	142,78
0,2	136,10	248,12	165,30	197,13	134,48	160,97	140,18
0,3	**135,87**	229,75	160,55	182,05	**133,89**	**160,64**	139,03
0,4	136,22	217,63	157,88	172,84	133,93	160,99	**138,62**
0,5	137,03	208,94	156,57	166,91	134,43	161,88	138.76
0,6	138,25	202,47	**156,34**	163,24	135,34	163,28	139,35
0,7	139,90	197,61	157,16	**161,46**	136,67	165,23	140,42
0,8	142,06	194,07	159,20	161,62	138,48	167,89	142,03
0,9	144,96	191,77	163,16	164,57	140,95	171,60	144,42
1,0	149,04	**191,18**	174,16	179,47	144,94	177	148,43

TAB. IV.2 – Comparaison de perplexités avec et sans un ML thématique

5.3 Résultats

Les résultats des évaluations de perplexités sont donnés dans la table IV.2 et la figure IV.2. Il est intéressant de rapprocher la table IV.1, concernant la taille du corpus d'apprentissage, et la table IV.2, indiquant la valeur du coefficient λ à affecter au modèle général. Par exemple, le corpus de *Sport*, dont la taille du corpus d'apprentissage est très faible, a un modèle interpolé avec un coefficient λ de 0,7 pour le modèle général, alors que les thèmes *Etranger* et *Culture*, dont les données d'apprentissage sont de taille suffisante, ont un coefficient λ à 0,3. D'une façon globale, on constate que le coefficient λ, appliqué au modèle général, reflète la qualité du bigramme thématique. De plus, il est intéressant de remarquer que les perplexités des modèles thématiques de *Etranger, Economie, Culture et Politique*, dont les données d'apprentissage suffisent, sont meilleures que celles du modèle général.

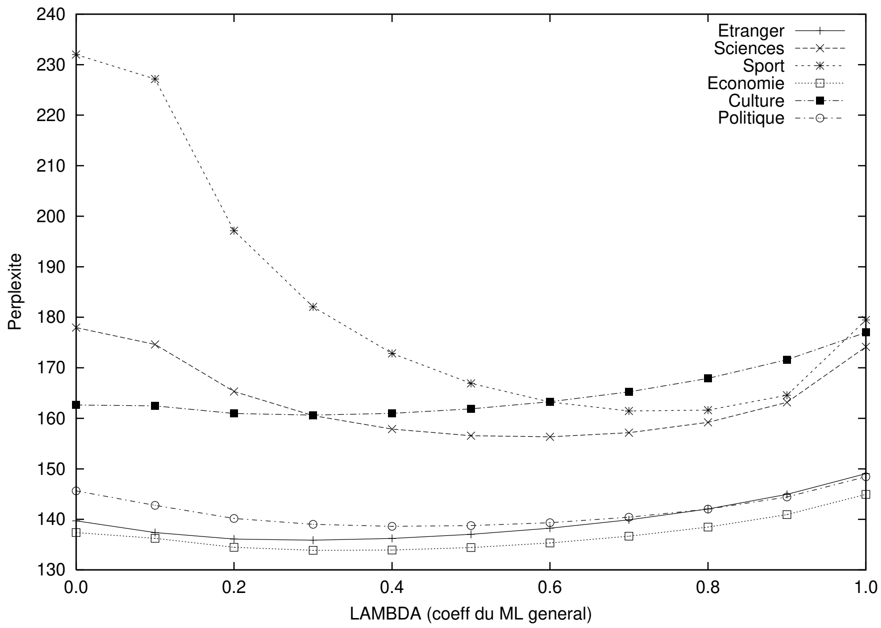


FIG. IV.2 – Evolution des perplexités des modèles bigrammes combinés

Dans tous les cas, sauf *Histoire*, le modèle interpolé permet de gagner en perplexité. En effet, ce thème est une exception car le corpus d'apprentissage est relativement faible et ses articles portent souvent sur des sujets relatifs aux thèmes *Culture, Potilique ou Etranger*. Une synthèse des résultats signifiant les gains des modèles combinés est présentée en table IV.3. Un gain moyen de l'ordre de 8,7 % est observé sur les 6 thèmes concernés.

	λ	PP ML combiné	PP ML gral	Gain
Etranger	0,3	135,87	149,04	8,84 %
Histoire	1	-	191,18	-
Science	0,6	156,34	174,16	10,23 %
Sport	0,7	161,46	179,47	10,04 %
Economie	0,3	133,89	144,94	7,62 %
Culture	0,3	160,64	177	9,24 %
Politique	0,4	138,62	148,43	6,61 %

TAB. IV.3 – Perplexités de modèles bigrammes combinés

6 Perspectives

Dans ce chapitre, nous avons abordé le problème de l'adaptation des modèles de langage. Nous nous sommes intéressés aux solutions proposées dans la littérature pour l'évaluation des paramètres des modèles, ainsi qu'à plusieurs modèles existants. Notre objectif était de valider l'idée qu'un modèle qui intègre des connaissances thématiques sera plus performant qu'un modèle généraliste, et ce, en déterminant automatiquement le thème du texte. Les expérimentations que nous avons conduites nous ont permis d'arriver à cette conclusion, avec un gain de perplexité de 8,7 %, en comparant des modèles de langage bigrammes thématiques interpolés et un modèle généraliste.

Pour aller encore plus loin dans cette voie, il est nécessaire de disposer d'un ensemble de corpora d'apprentissage thématiques qui permettront de disposer de modèles de langage plus précis, par exemple de type trigrammes. Dans le chapitre V, nous développons des méthodes de segmentation automatique de documents qui peuvent être utilisées pour segmenter thématiquement un corpus pour l'apprentissage de nouveaux modèles.

Chapitre V

Segmentation thématique

1 Introduction

Segmenter une entité textuelle ou audio consiste à la découper en fragments de nature proche. Plusieurs types de segmentations sont possibles. On peut, par exemple, analyser les textes suivant leur structure physique (phrases, paragraphes, titres, marques typographiques...) ou selon leur contenu (styles d'écriture, liens thématiques entre les termes employés...). La *segmentation thématique* a pour but de déterminer automatiquement les changements de thèmes entre les segments qui composent un document.

Dans l'objectif d'améliorer les systèmes de reconnaissance automatique de la parole, la segmentation peut être utile pour séparer en plusieurs thèmes les corpora d'apprentissage des modèles. En recherche documentaire, la segmentation de documents peut être un atout, par exemple, pour répondre le plus précisément possible à la question posée par l'utilisateur. Ainsi, il disposera uniquement des segments thématiques demandés, et non du document complet. Dans ce cas, la segmentation a pour but de trouver les blocs thématiquement homogènes des documents : les unités documentaires du texte. La segmentation peut également être utilisée pour rapporter des documents entiers ; dans ce cas, elle permet la mise en valeur de certains documents grâce à un nouveau calcul des similarités avec la requête à partir du découpage des textes en unités documentaires.

Dans ce chapitre, nous introduisons de nouvelles méthodes pertinentes pour la segmentation thématique de documents. L'objectif est de retrouver les frontières de paragraphes adjacents de thèmes différents. Pour effectuer les évaluations, nous avons créé des corpora de test dans lesquels chaque document est constitué de 3 paragraphes tirés aléatoirement dans nos corpora thématiques. Pour mesurer l'impact que peut avoir la taille des paragraphes qui composent les documents sur la qualité de la segmentation, nous avons défini trois corpora dont les contraintes de tailles sont différentes. Un premier corpus de 355 documents n'est constitué que de paragraphes supérieurs à 300 entités lexicales (appelé "Sup300"). Un deuxième corpus de 820 documents n'est constitué que de paragraphes de taille inférieure à 300 entités lexicales (appelé "Inf300"). Un troisième corpus de 1393 documents ne prend en compte aucune contrainte de taille (appelé "Test"). Toutes les évaluations sont effectuées sur ces 3 corpora mais les résultats présentés graphiquement concernent essentiellement le corpus de test.

La segmentation thématique que nous proposons opère en deux étapes. Nous procédons, dans un premier temps (section 2), à un repérage de ruptures candidates, puis à leur sélection (section 3). Plusieurs méthodes ont été développées pour chacune de ces étapes. Afin de les valider, nous évaluons le taux de détection des ruptures thématiques avec les valeurs de *rappel* et *précision* suivantes :

$$rappel = \frac{nombre\ de\ ruptures\ correctement\ détectées}{nombre\ de\ ruptures\ à\ trouver} \tag{V.1}$$

$$précision = \frac{nombre\ de\ ruptures\ correctement\ détectées}{nombre\ de\ ruptures\ totales\ détectées} \tag{V.2}$$

En principe, une rupture thématique ne peut pas se trouver dans le courant d'une phrase. C'est pourquoi, dans l'ensemble des traitements, nous ne chercherons les ruptures thématiques qu'en fin de phrase.

2 Repérage des candidats

2.1 Modèle à base de mémoire cache

Comme décrit dans le chapitre III, le modèle à base de mémoire cache est un outil performant pour la classification thématique, c'est pourquoi nous avons voulu l'utiliser dans le cadre de la segmentation. Le contenu de la mémoire cache est une représentation de l'historique puisqu'elle en conserve uniquement les N derniers mot clés. La mémoire cache est réinitialisée à chaque nouveau document. Les probabilités du modèle sont calculées à partir des distances $d_j^*(i)$, distances de Kullback-Liebler normalisées évaluées entre le contenu de la mémoire cache et les histogrammes de mots clés de chaque thème T_j. Pour la classification thématique, le thème assigné au paragraphe est celui dont la distance est la plus petite.

Les documents sont segmentés en phrases, et la distance de Kullback-Liebler est calculée à chaque fin de phrase. La distance $d_j^*(i)$ est la distance entre la mémoire cache et l'histogramme du thème T_j à la fin de la i-ème phrase du document. Nous nous intéressons à l'évolution de cette distance pour le meilleur thème, au sens de la classification. Cette variation s'exprime par :

$$\delta(i) = d_j^*(i) - d_j^*(i-1)$$

où j est le meilleur thème à la fin de la $(i-1)$ème phrase. Nous proposons une rupture candidate à chaque augmentation importante de la distance, c'est à dire quand :

$$\delta(i) > \theta$$

où θ est un seuil déterminé expérimentalement. Ceci correspond à une chute de la probabilité du meilleur thème, comme on peut l'observer sur la figure V.1. Cette figure montre un exemple d'évolution des probabilités du modèle cache lorsque deux paragraphes de thèmes différents se succèdent. Elle permet de constater que la probabilité du thème du premier paragraphe décroit fortement dès que le second paragraphe est abordé. Cependant, il n'y a pas de changement "radical" du thème, car lorsque le second paragraphe est traité, le cache contient les données du premier paragraphe (100 mots clés).

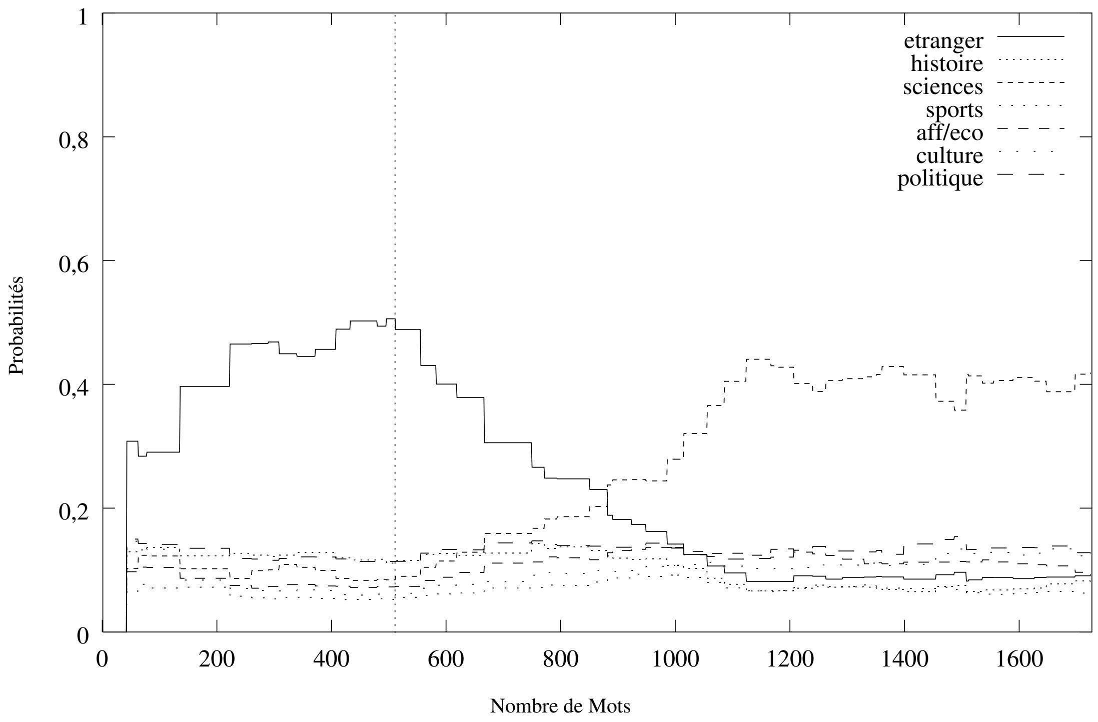


FIG. V.1 – Exemple de classification thématique d'un document composé de deux paragraphes : le premier est extrait du thème Etranger, et le second (qui commence au 515ème mot) provient du thème Sciences.

Les résultats que l'on obtient avec cette méthode sont exprimés dans la table V.1 et la figure V.2. On observe une valeur élevée de rappel, ce qui signifie que peu de frontières thématiques n'ont pas été détectées. Par contre, la faible valeur de précision indique un nombre important de fausses alarmes. Ce phénomène est moins marqué sur le corpus dont les paragraphes sont inférieurs à 300 mots car un nombre plus élevé de ruptures candidates y est plus adapté. Les différentes valeurs de θ ne font que faire varier proportionnellement les valeurs de rappel et de précision.

θ		0,001	0,0012	0,0014	0,0016	0,0018	0,002	0,003	0,004
Sup300	Rappel	0,9842	0,9827	0,9655	0,9496	0,9281	0,8978	0,6691	0,4331
	Précision	0,0975	0,119	0,1403	0,1662	0,1907	0,2197	0,3279	0,3859
Inf300	Rappel	0,8244	0,7939	0,7706	0,731	0,7064	0,6682	0,5185	0,3966
	Précision	0,3921	0,4108	0,4362	0,453	0,477	0,4952	0,5869	0,6581
Test	Rappel	0,9091	0,8896	0,8664	0,8444	0,8152	0,789	0,612	0,4332
	Précision	0,1261	0,1495	0,1747	0,2018	0,2309	0,2608	0,4047	0,4798

TAB. V.1 – Résultats de la segmentation par la méthode de repérage des ruptures candidates qui utilise le modèle à base de mémoire cache.

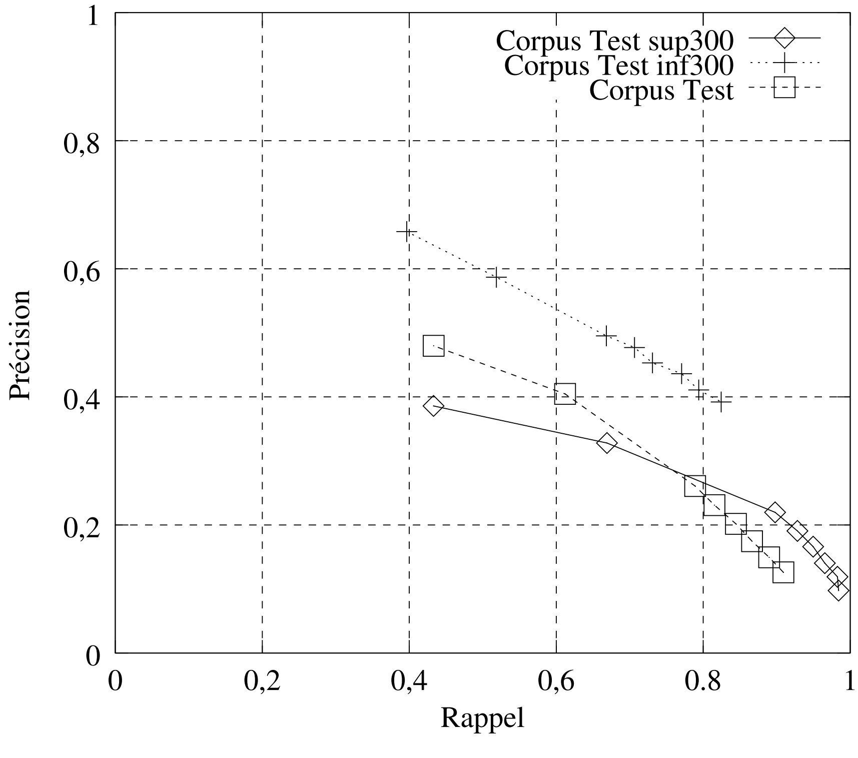


FIG. V.2 – Résultats de la segmentation des 3 corpora par la méthode de repérage des ruptures candidates qui utilise le modèle à base de mémoire cache.

2.2 Bigrammes à distance

On appelle un "cache arrière" ou "backward cache" des bigrammes à distance. Nous avons effectué un apprentissage de bigrammes en utilisant la mémoire cache qui précède chaque mot clé w_i. Ainsi, pour chaque mot clé du texte, on considère toutes les paires de mots clés composées de w_i avec chacun des mots clés que l'on observe dans sa mémoire cache. Ce processus itère selon le schéma de la figure V.3.

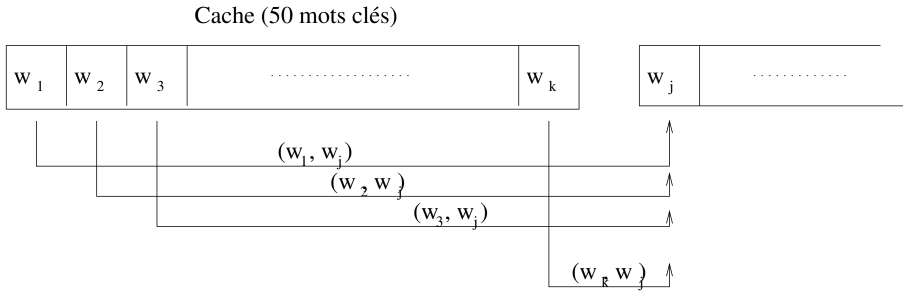


FIG. V.3 – Méthode d'apprentissage du backward cache

Pour respecter les contraintes de mémoire, nous avons sélectionné seulement les 4000 mots clés les plus probables, parmi les 11575 mots clés des 7 thèmes, et nous avons limité la taille du cache à 50 mots clés.

Dans le cadre de la segmentation thématique, nous cherchons à évaluer si une rupture est présente avant la i-ème phrase. Pour ce faire, nous disposons de S_i, la phrase qui suit i, et de C_i, la mémoire cache qui précède i. Ceci est présenté dans la figure V.4.

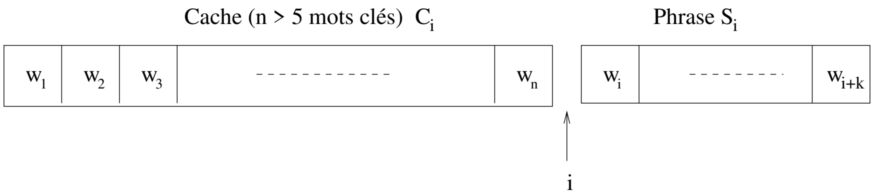


FIG. V.4 – Méthode d'évaluation du contexte d'un backward cache (bigrammes à distance) : on cherche à déterminer si i est une rupture thématique.

L'objectif est de définir la quantité d'information que C_i et S_i ont en commun afin d'en déduire si i est une frontière thématique. Pour effectuer cette évaluation, nous utilisons $P_c(w_i\ C_i)$, la probabilité d'avoir une continuité à l'instant i, et $P_r(w_i\ C_i)$, la probabilité d'avoir une rupture. Ceci s'évalue :

$$P_c(w_i\ C_i) = \sum_{n \in C_i} P(w_i, w_n)$$

et

$$P_r(w_i\ C_i) = \sum_{n \in C_i} P(w_i)P(w_n)$$

où :

$$P(w_i, w_n) = \frac{n(w_i, w_n)}{\sum_{k \in C_i} n(w_i, w_k)}$$

avec $n(w_i, w_n)$ est le nombre de fois où w_n a fait partie de la mémoire cache de w_i pendant la phase d'apprentissage. P_c représente le fait que w_i et w_n partagent beaucoup d'information et P_r exprime leur indépendance l'un envers l'autre.

Pour déterminer s'il y a une rupture thématique, nous définissons un rapport normalisé $R(i)$, calculé pour chaque mot w_i de S_i tel que :

$$R(i) = \frac{1}{1 + \frac{1}{R_i'(w_i)}}$$

$$R'(i) = \frac{P_c(w_i\ C_i)}{P_r(w_i\ C_i)}$$

Dans ce cas, i est un candidat à la rupture quand le plus petit des $R(i)$ est inférieur à un seuil :

$$\min_{k \in S_i}(R(i+k)) < \delta \tag{V.3}$$

Il faut noter que $R(i)$ varie entre 0 et 1.

Les résultats obtenus par cette approche sont présentés dans la table V.2. Comme on pouvait s'y attendre, les résultats obtenus ne sont pas à la hauteur de ceux rencontrés avec le modèle à base de mémoire cache. Ceci s'explique par le fait que, contrairement au modèle cache, dans le cas présent on cherche les frontières thématiques en n'ayant aucune connaissance concernant les thèmes. Concernant les différences entre les corpora, il apparaît que cette méthode est plus efficace sur des paragraphes de taille suffisemment importante. Ceci est dû au fait qu'il faut que C_i soit de taille suffisante et de même thème pour être cohérente.

Sup300	Rappel	0,9151
	Précision	0,1332
Inf300	Rappel	0,4128
	Précision	0,4953
Test	Rappel	0,7235
	Précision	0,1494

TAB. V.2 – Résultats de la segmentation par la méthode de repérage de bigrammes à distance.

2.3 Synthèse des résultats du repérage

A fin de comparaison, nous avons aussi utilisé une méthode systématique. Elle place arbitrairement un candidat toutes les N phrases. Cette méthode permet avec $N = 1$ de trouver toutes les ruptures (rappel=1), avec la valeur minimale de précision qui peut être observée. Les résultats, dans ce cas sont dans le tableau V.3.

N		1	5	10
Sup300	Rappel	1	1	0,9913
	Précision	0,0178	0,0907	0,1838
Inf300	Rappel	1	0,6792	0,3629
	Précision	0,1551	0,6272	0,8989
Test	Rappel	1	0,8391	0,6816
	Précision	0,0249	0,1084	0,1805

TAB. V.3 – Résultats de la segmentation si l'on place un candidat toutes les N phrases.

La figure V.5 synthétise les résultats obtenus sur le corpus de test par les 3 méthodes de repérage présentées dans cette section. Signifier des ruptures thématiques sans disposer de connaissances thématiques semble être une tâche ardue. Ainsi, le backward cache donne des résultats quasiment identiques à ceux de la méthode systématique, cependant, ce modèle pourrait être amélioré si le nombre de mots clés pris en compte était plus élevé. Le modèle à base de mémoire cache (dépendant des thèmes appris) reste la meilleure solution que nous avons testée.

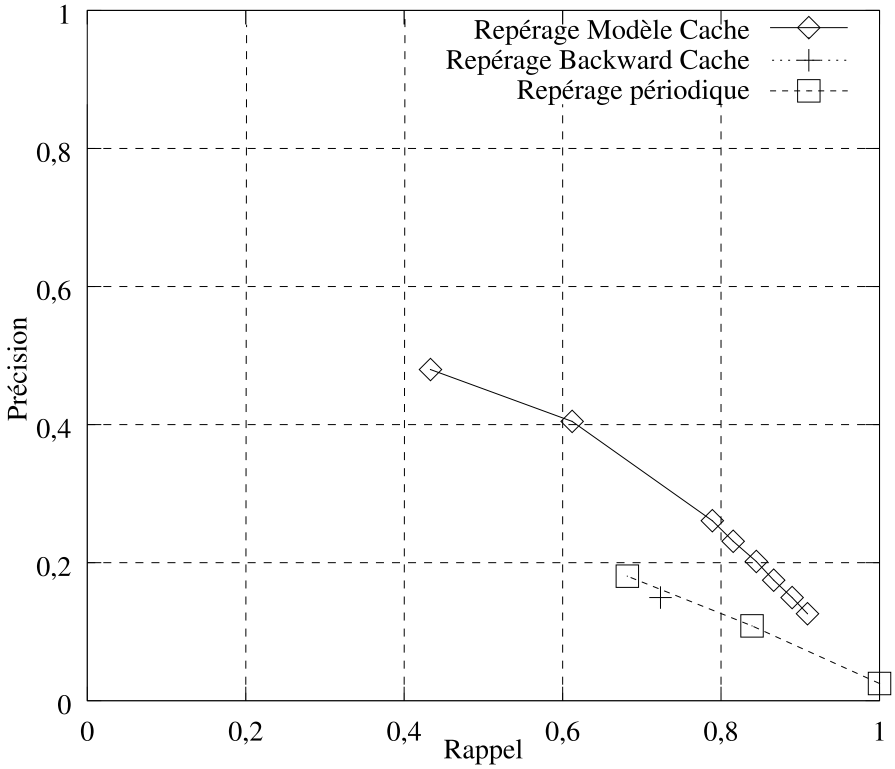


FIG. V.5 – Résultats de la segmentation sur le corpus de test avec 3 méthodes de repérage des ruptures candidates.

3 Sélection des candidats

Dans cette section, nous définissons un segment comme étant la portion de texte comprise entre deux ruptures candidates. Dans une première étape, nous avons exposé des méthodes pour repérer un ensemble de ruptures candidates qui peuvent représenter la frontière thématique entre deux segments de texte. Nous avons pu observer de forts taux de rappel. Dans cette seconde étape, nous développons plusieurs méthodes dont l'objectif est de sélectionner les candidats en minimisant les fausses alarmes afin de gagner en valeur de précision, sans trop perdre de la valeur de rappel.

3.1 Modèle cache

En utilisant le modèle à base de mémoire cache pour leur classification, il est possible que deux segments successifs soient labéllisés avec le même thème. Les ruptures de thèmes sont alors définies lorsque deux labels thématiques différents sont observés dans deux segments adjacents. On peut noter que, comme les ruptures candidates sont obtenues avec des règles locales appliquées au contenu de la mémoire cache, quand on génère des segments de textes qui ne contiennent pas un nombre suffisant de mots, la rupture candidate est alors ignorée. Ce nombre a été fixé empiriquement en fonction de la méthode de repérage utilisée.

Les résultats sont dans la table V.4 en fonction de la méthode de détection des candidats. On observe des résultats nettement meilleurs sur le corpus Sup300 que sur les autres corpora car sur ces derniers, la mémoire cache ne contient pas toujours un nombre de mots suffisant pour être représentatif. Ces résultats sur le corpus "Test" sont en figure V.6. On constate à nouveau que la solution entièment basée sur le modèle cache est plus performante que les autres méthodes. De plus, on remarque l'importance de la méthode de repérage lors de la phase de sélection, puisque lorsque les candidats sont choisis par la méthode à base de mémoire cache on obtient de meilleurs résultats qu'avec la méthode systématique. On le voit notamment sur le corpus "Test" en comparant les cas où $\theta = 0,0018$ et $N = 10$. Pour la même valeur de rappel, la valeur de précision est nettement meilleure par le repérage à base de mémoire cache. Le problème de cette méthode est qu'elle requiert des connaissances *a priori* sur les thèmes. C'est pourquoi, nous avons orienté nos travaux vers une nouvelle méthode de sélection, indépendante des thèmes.

θ		0,001	0,0012	0,0014	0,0016	0,0018	0,002	0,003	0,004
Sup300	Rappel	0,8317	0,8273	0,8086	*0,7813*	0,754	0,7281	0,5194	0,3367
	Précision	0,6856	0,697	0,7168	*0,7479*	0,7594	0,7749	0,8186	0,8357
Inf300	Rappel	0,0376	0,0421	0,0467	0,0499	0,0538	0,0577	0,059	0,0544
	Précision	0,8169	0,8333	0,8605	0,8556	0,8557	0,8476	0,8585	0,875
Test	Rappel	0,3857	0,3868	0,3797	0,3775	0,3715	0,3618	0,2851	0,2035
	Précision	0,6492	0,6675	0,6794	0,7212	0,7394	0,7543	0,8141	0,85

Méthode de repérage		Backward cache	Périodique		
			N=1	N=5	N=10
Sup 300	Rappel	0,6748	0,6475	0,8144	0,6906
	Précision	0,5563	0,3719	0,4825	0,5184
Inf 300	Rappel	0,0635	0,2029	0,2139	0,0966
	Précision	0,8448	0,5217	0,6613	0,8418
Test	Rappel	0,3419	0,3958	0,4706	0,3726
	Précision	0,5434	0,4235	0,4803	0,4911

TAB. V.4 – Sélection des candidats par le modèle cache

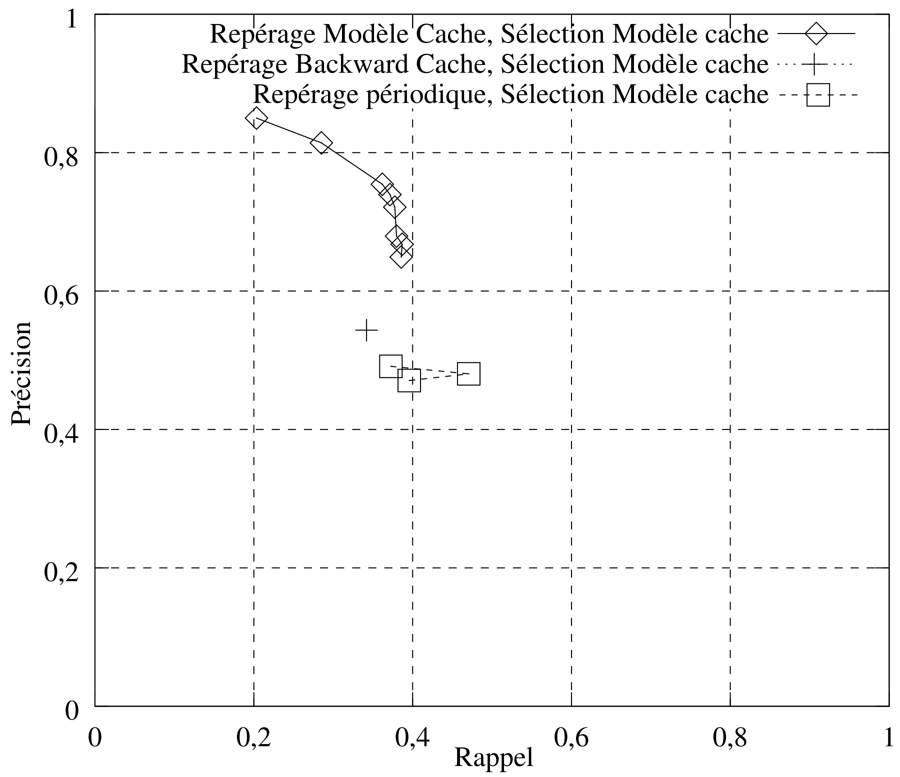


FIG. V.6 – Selection des candidats avec le modèle cache

3.2 Historique variable

Le principe global de cette approche consiste à déterminer la séquence optimale de ruptures parmi l'ensemble des ruptures candidates. L'idée est de calculer la distance entre le segment passé et le segment à venir. On déduit si une rupture est réelle ou si c'est une fausse alarme en faisant varier la taille des historiques possibles d'une rupture candidate. Ceci revient à se poser les questions suivantes : Est-ce que les segments précédents formaient une "entité thématique" et est-ce que celle-ci porte sur le même thème que le prochain segment ? On note r, le cas où la rupture candidate est effective, et c le cas où c'est une continuité. Cette approche est une programmation dynamique dont le fonctionnement général est représenté par l'automate de la figure V.7.

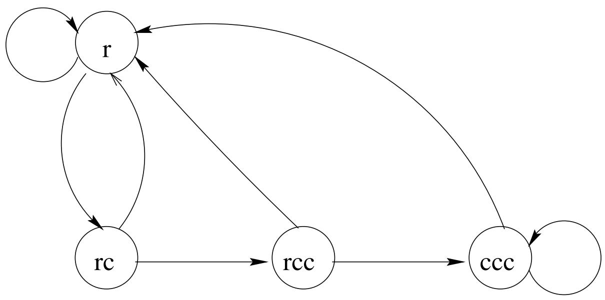


FIG. V.7 – Automate de la méthode de sélection des candidats

Le treillis d'évaluation qui résulte de cet automate est présenté dans la figure V.8. On note RC_i la rupture candidate au i-ème segment.

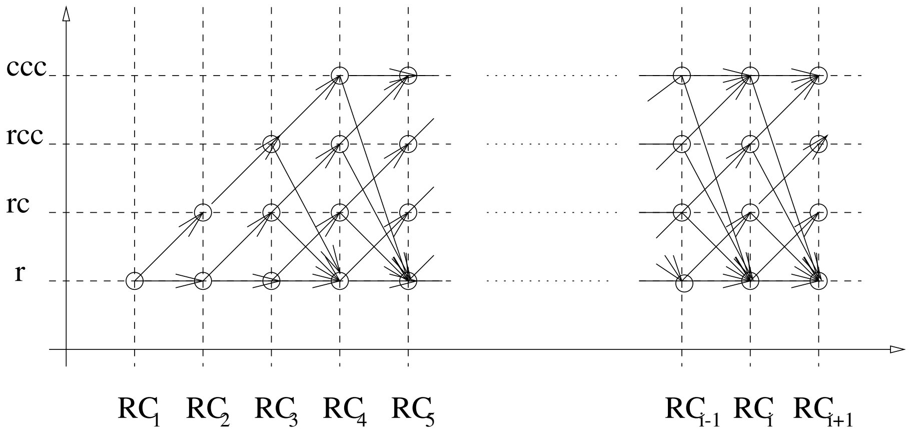
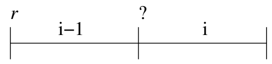


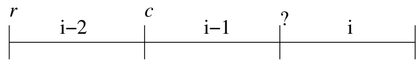
FIG. V.8 – Treillis de la programmation dynamique selon l'automate de la figure V.7

L'évaluation des points de ce treillis s'effectue en 3 étapes. Dans un premier temps, on évalue la distance $d_i(G, D)$, distance de Kullback-Liebler entre le contexte gauche et le contexte droit, du i-ème candidat, où un nombre différent de segments peut être utilisé pour représenter le contexte gauche, en fonction de l'état pour lequel on calcule cette distance, comme suit :

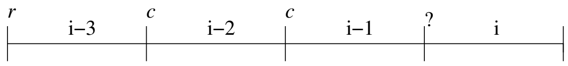
– état r : Le contexte gauche est représenté par un seul segment (i. e. la rupture candidate au $(i-1)$-ème segment est une rupture effective).



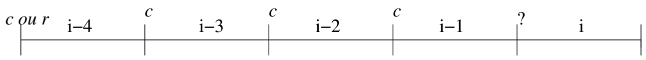
– état rc : Le contexte gauche contient les deux segments précédents.



– état rcc : Le contexte gauche contient les trois segments précédents.



– état 4 ccc : Le contexte gauche contient les quatre segments précédents.



Ensuite, on analyse la variation de cette distance avec celle du même état précédent :

$$\Delta_i = d_i(G_i, D_i) - d_{i-1}(G_{i-1}, D_{i-1})$$

Cette distance permet de calculer la probabilité de continuité $P(c)$ telle que :

$$P(c) = \frac{\alpha}{1 + \exp^{-\Delta_i}}$$

et celle d'une rupture $P(r)$ telle que : $P(r) = 1 - P(c)$. L'état précédent que l'on choisit est celui dont $P(c)$ est la plus grande. On remarque que dans le cas où $\Delta_i = 0$ et $\alpha = 1$, on aura $P(c) = P(r) = 0,5$. Les résultats de cette méthode sont en table V.5 et en figure V.9 pour le corpus "Test".

Δ		0,001	0,002	0,002	0,003
α		3	2	3	3
Test sup 300	Rappel	0,895	0,7050	0,6547	0,3813
	Précision	0,3024	0,4545	0,5078	0,5591
Test inf 300	Rappel	0,5159	0,3513	0,3454	0,2048
	Précision	0,688	0,6691	0,6886	0,6695
Test	Rappel	0,7112	0,5634	0,5477	0,3423
	Précision	0,3394	0,4906	0,5502	0,6417

Methode de repérage		Backward cache	Périodique			
			5	5	5	10
α		3	1	2	3	1
Sup300	Rappel	0,7252	0,9727	0,9079	0,8086	0,823
	Précision	0,3871	0,1798	0,2965	0,3949	0,324
Inf300	Rappel	0,1607	0,2793	0,2735	0,2722	0,0305
	Précision	0,5636	0,6274	0,6604	0,6699	0,7231
Test	Rappel	0,4867	0,6012	0,5705	0,5447	0,4284
	Précision	0,3524	0,1587	0,2607	0,3539	0,2473

TAB. V.5 – Sélection des candidats par programmation dynamique

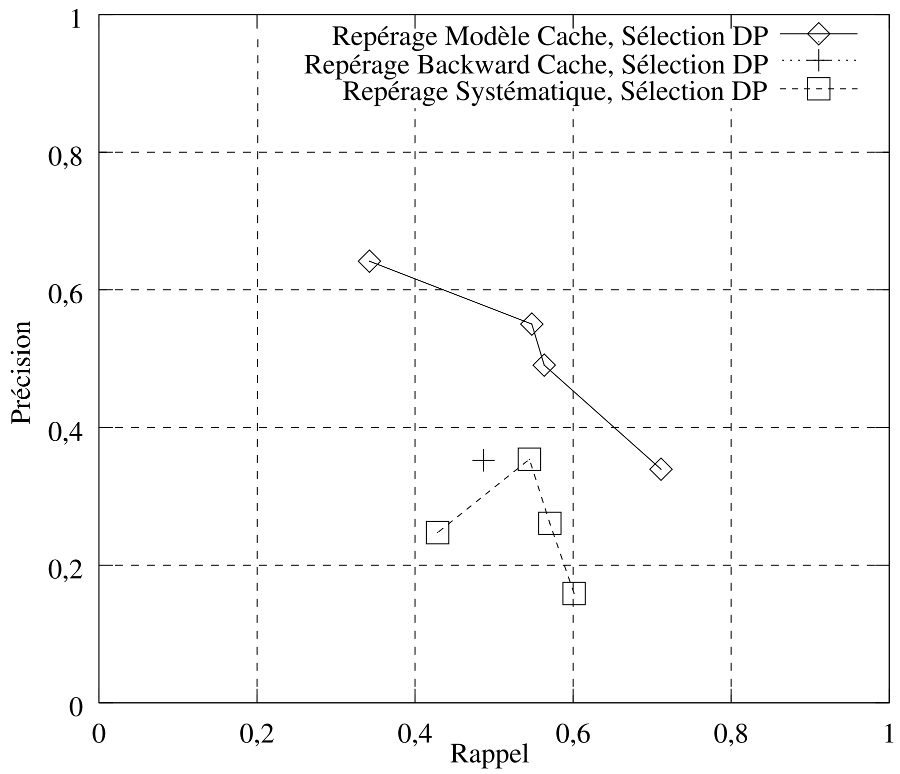


FIG. V.9 – Sélection des candidats avec un historique variable.

4 Synthèse des résultats

Ce chapitre a été consacré à un ensemble de méthodes dont l'objectif est la segmentation thématique de documents. La solution que nous proposons sépare le problème en deux tâches : le repérage de ruptures candidates et leur sélection. Deux méthodes sont décrites pour chacune de ces tâches. Pour chacune des deux tâches, la première méthode s'appuie sur le modèle à base de mémoire cache, outil de classification, et la seconde méthode est une tentative de segmentation sans connaissances thématiques. La figure V.10 donne la courbe de rappel et précision que l'on obtient avec l'ensemble des différentes méthodes utilisées. Ces résultats montrent que différentes stratégies doivent être utilisées selon les valeurs de rappel ou de précision que l'on cherche à obtenir. L'ensemble des meilleurs points de ces courbes sont trouvées avec le modèle cache, car il est difficile de segmenter thématiquement un document sans connaissances thématiques préalables.

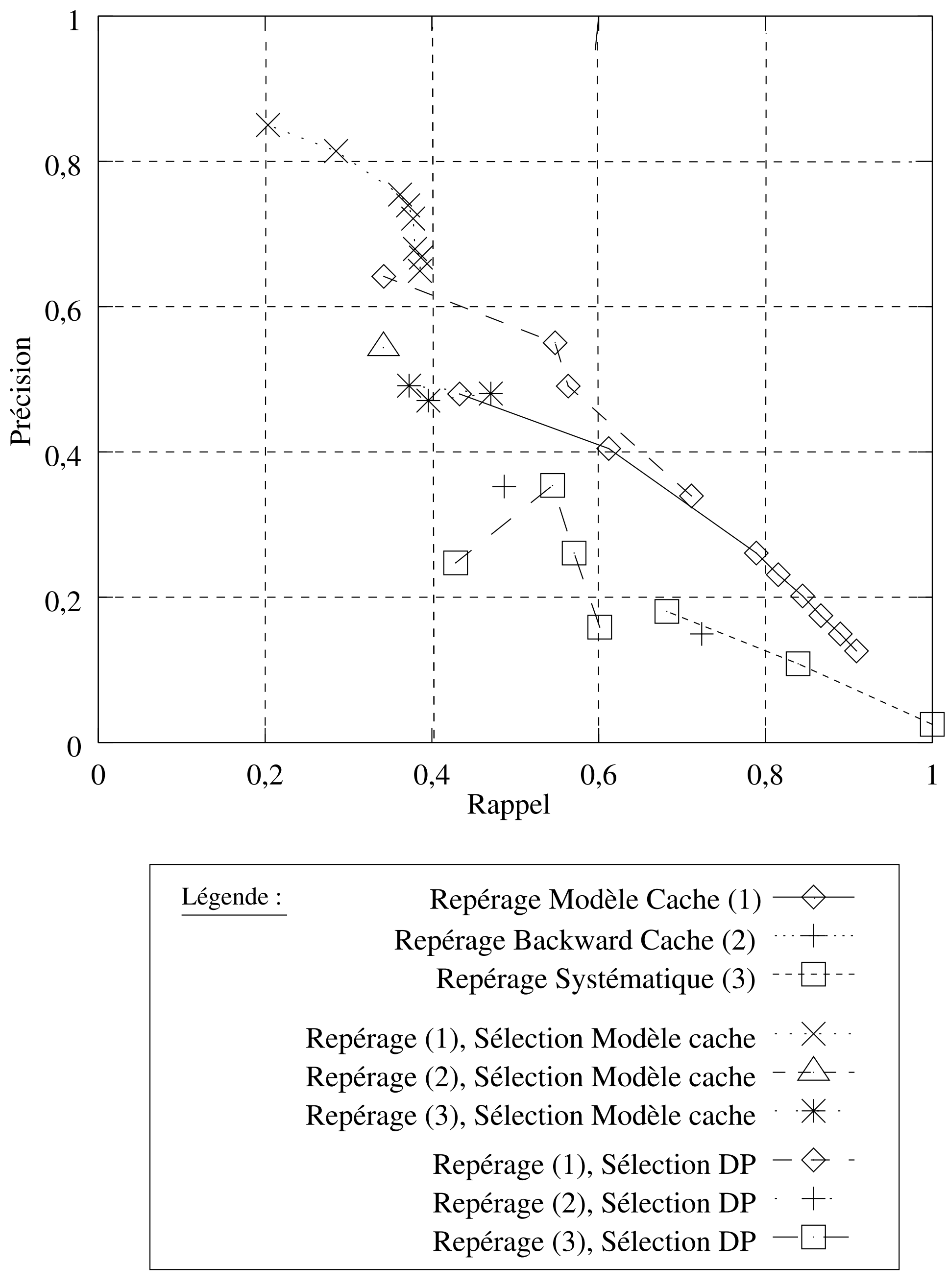


FIG. V.10 – Synthèse des résultats de segmentation thématique

Chapitre VI

Expansion de requête

1 Introduction

Les systèmes de recherche documentaire prennent en entrée la question d'un utilisateur et y répondent en proposant un ensemble de documents, extrait d'une banque de données qui correspondent au mieux à la question posée. Un des problèmes de ces systèmes vient du fait que l'utilisateur n'a aucune connaissance de la structure de la banque de données. Par ailleurs, selon que l'utilisateur est familier des systèmes de recherche ou non, la formulation de la question ainsi que les réponses attendues peuvent varier énormément.

Le volume croissant d'informations pose également de nombreux problèmes en recherche documentaire. La quantité excessive des résultats retournés fait que les utilisateurs n'examinent pas toutes les différentes pages de réponses. Ainsi le défi est de concevoir un système de recherche qui proposera une bonne précision tout en conservant une valeur suffisante de rappel pour satisfaire la majorité des utilisateurs. De plus, l'ordre dans lequel les réponses sont proposées est essentiel.

L'une des causes d'échec des systèmes de recherche documentaire vient de la non concordance entre le vocabulaire de la question posée et celui des documents. On peut également ajouter l'ambiguïté inhérente des mots du langage naturel. Quand un document pertinent ne contient pas les termes qui sont dans la requête, ce document ne sera pas retourné par le système. Le but de l'expansion de requête est de réduire ce manque de correspondance entre le vocabulaire de la question et celui du document en majorant la requête avec des mots ou des expressions dont la signification est semblable ou s'il y a une autre relation d'ordre statistique entre la question et le document. Ce procédé peut prendre plus d'importance encore lorsque la recherche documentaire est appliquée à des documents de parole, puisque les erreurs de correspondance du vocabulaire sont accentuées par la présence d'erreurs dans la transcription automatique des documents parlés.

Cette modification de la requête initiale augmente le nombre de termes de la question en fonction d'un modèle fondamental de recherche. Elle peut être faite manuellement (par exemple [Brajnik et al. 1996]) ou construite automatiquement ([Carpineto et Romano 1998], [Cooper et Byrd 1997]) en amenant l'utilisateur à reformuler sa question, en mode interactif. Une autre possibilité, plus élaborée, est d'exploiter les rapports que l'on peut observer entre le contenu des documents dans une collection. Les systèmes qui utilisent ce type de méthode sont gourmands en calculs et n'ont pas montré jusqu'ici des avantages réels par rapport aux systèmes qui utilisent la requête initiale. Une troisième approche pour limiter les problèmes de vocabulaire consiste à extraire automatiquement certains termes les plus courants parmi les documents les plus pertinents fournis par la requête initiale. L'intérêt croissant de cette technique met en avant le besoin de développer des méthodologies bien fondées pour le classement des termes de l'expansion.

La section 2 présente les approches de base utilisées en expansion de requête. Cependant, les nombreuses variantes présentes dans la littérature ne permettent pas un énoncé exaustif. Dans la section 3, on introduit une nouvelle fonction d'évaluation des termes qui est basée sur les différences entre la distribution des meilleurs termes des documents obtenus avec la requête initiale et la distribution des termes dans tous les documents. On propose une méthode de calcul simple pour assigner des scores aux termes candidats à l'expansion de la requête. La méthode est basée sur la mesure de divergence de Kullback-Leibler. Cette solution est la transposition au domaine de l'expansion de requête du modèle à base de mémoire cache présenté en classification thématique (chapitre IV). Ce travail est issu d'une collaboration avec la "*Fondazione Ugo Bordoni*" (Rome, Italie) avec M. Claudio Carpineto et M. Giovanni Romano ([Carpineto et al. 1998]). En section 4, on propose les résultats obtenus sur les données des campagnes TREC-7 et TREC-8.

2 Approches en expansion automatique de requête

La plupart des systèmes d'expansion de requête utilisent la version améliorée de la formule initiale de Rocchio ([Rocchio 1971], [Salton et Buckley 1990]). Elle est le point de départ pour mettre à jour les poids de termes :

$$Q_{new} = \alpha Q_{orig} + \frac{\beta}{|R|} \sum_{r \in R} r - \frac{\gamma}{|R'|} \sum_{r' \in R'} r' \tag{VI.1}$$

où Q_{new} est un vecteur de poids des termes pour la requête augmentée, Q_{orig} est un vecteur de poids des termes pour la requête initiale, R et R' sont respectivement les ensembles des documents pertinents et non-pertinents, r et r' sont deux vecteurs de poids de termes extraits respectivement à partir de R et de R'. Les poids dans chaque vecteur sont calculés à l'aide de la collection entière.

Si le classement des termes et l'expansion de requête s'appuient sur un ensemble de documents dont les scores sont élevés, la formule VI.1 se réduit à :

$$Q_{new} = \alpha Q_{orig} + \frac{\beta}{|R|} \sum_{r \in R} r \tag{VI.2}$$

où R est l'ensemble des meilleurs documents fournis par le système comme étant pertinents. Les poids obtenus par cette formule sont typiquement utilisés pour l'ordonnancement et la ré-évaluation des poids des termes de la question. Cette approche est simple et efficace, mais elle a l'inconvénient que chaque poids d'un terme reflète l'utilisation de ce terme en rapport avec la collection entière, indépendamment de la requête formulée.

Une approche conceptuellement différente pour évaluer la convenance d'un terme est basée sur l'analyse de distribution. Afin de distinguer les bons termes pour l'expansion et les termes de moindre intérêt, il peut être plus pratique de comparer l'occurrence dans les documents appropriés avec celle dans tous les documents, pour chaque requête donnée. En d'autres termes, on peut supposer que la différence entre la distribution des termes dans la collection globale de document et la distribution des mêmes termes dans un ensemble de documents appropriés est un bon indicateur d'une différence sémantique. En particulier, on s'attend à ce que la fréquence des termes appropriés soit plus élevée dans les documents appropriés que dans la collection entière, alors que d'autres termes apparaitront avec la même fréquence (aléatoirement) dans les deux ensembles de documents.

Une des premières applications de ce concept peut être trouvée dans ([Doszcocks 1978]), où une analyse statistique comparative des occurrences de termes - par l'intermédiaire d'une variante chi-carrée - a été établie pour suggérer des termes potentiellement appropriés pour l'expansion interactive de requête. Dans ([Robertson 1990]), on présente un cadre théorique plus général qui supporte l'utilisation des différences dans la distribution de termes afin de choisir et donner un poids aux termes à inclure dans la requête augmentée. Il est montré que, l'inclusion du terme t dans la requête augmentée accroît l'efficacité de recherche par $w_t(p_t - q_t)$, où w_t est le poids original du terme t, p_t est la probabilité qu'un document approprié contienne le terme t, et q_t est la probabilité qu'un document non-approprié contienne le terme t.

Des variantes de l'ordonnancement des termes proposées dans [Robertson 1990] ont été utilisées ultérieurement dans divers systèmes avec différentes fonctions de calcul des poids et différentes méthodes pour estimer p_t et q_t ([Buckley et Salton 1995], [Robertson et al. 1995], [Hawking et al. 1998]). Dans [Efthimiadis 1993], l'évaluation de la capacité à ranger des termes potentiellement appropriés est effectuée avec plusieurs fonctions d'obtention des termes basées sur l'analyse de distributions. Cette évaluation a été confinée à un scénario de recherche interactive et n'a pas porté sur l'expansion automatique de requête.

3 Expansion de requête automatique par retour de pertinence

Le modèle d'expansion de requête proposé dans cette section a pour caractéristique commune avec d'autres modèles généralement utilisés, de représenter un document ou une requête par des vecteurs de poids dans lesquels chaque poids correspond à un mot clé appelé *terme* appartenant à un vocabulaire V. On considère ces vecteurs comme des *vecteurs de termes*. Dans ce modèle, une question est représentée sous la forme d'un vecteur de termes Q, alors qu'un document est représenté par un vecteur de termes D. On suppose qu'une distance est définie dans l'espace de vecteur de termes et que, étant donnée une requête q, représentée par le vecteur Q, tous les documents concernant la requête sont représentés par des vecteurs appartenant à un positionnement $PD(q)$ tel que la distance entre n'importe quelle paire de vecteurs de $PD(q)$ (distance intraset) est inférieure à la distance entre n'importe quel vecteur de $PD(q)$ et n'importe quel vecteur extérieur $PD(q)$ (distance interset). Ces hypothèses peuvent être incertaines, mais elles sont communes à beaucoup de systèmes de recherche documentaire qui sont considérés, en pratique, parmi les meilleurs. Le problème de la recherche devient alors d'obtenir $PD(q)$ étant donné Q.

Malheureusement, particulièrement pour des requêtes courtes, la plupart des valeurs de Q sont égales à zéro, parce que la requête contient très peu de termes, alors que les vecteurs de $PD(q)$ ont, en général, beaucoup plus de valeurs différentes de zéro. En effet, les documents correspondants contiennent beaucoup plus de termes que la requête. Ceci rend la distance entre Q et les vecteurs de $PD(q)$ beaucoup plus élevée que beaucoup de distances interset. Ainsi, l'ensemble de vecteurs ayant la distance à partir de Q inférieure à un seuil donné est une approximation pauvre de $PD(q)$. Néanmoins, une meilleure approximation de $PD(q)$ a pu être trouvée, en considérant les vecteurs de termes comme étant limités à la distance d'un vecteur approprié Q_r.

Le modèle pour l'expansion de requête proposé dans cette section est basé sur une nouvelle proposition pour la détermination de Qr. On suppose que l'auteur de la requête q a eu l'intention de rechercher les documents représentés par des vecteurs de $PD(q)$. On note P_q, la distribution de probabilités unigrammes de tous les termes de V impliqués dans les vecteurs de $PD(q)$. Cette probabilité est inconnue et Q_r est construit à partir de son approximation. On note également P_C la distribution de probabilités unigrammes de tous les termes de V impliqués dans les vecteurs représentant les documents de la collection entière.

L'ensemble de vecteurs de termes $R(q)$ représentant les documents trouvés à partir de q par une méthode classique sans expansion de requête peut constituer un point de départ possible pour la construction de Q_r. P_R sera alors la distribution de probabilités unigrammes de tous les termes de V impliqués dans les vecteurs représentant les documents dans $R(q)$. Le vecteur Q_r aura comme éléments les fréquences de ces termes qui contribuent, la plupart du temps, à rendre P_R *divergent* par rapport à P_C. La mesure de divergence devra maximiser l'entropie relative entre D_R, l'ensemble des documents représentés dans $R(q)$, et D_C, l'ensemble des documents de la collection entière. Les différentes étapes de ce processus sont présentées dans la figure VI.1.

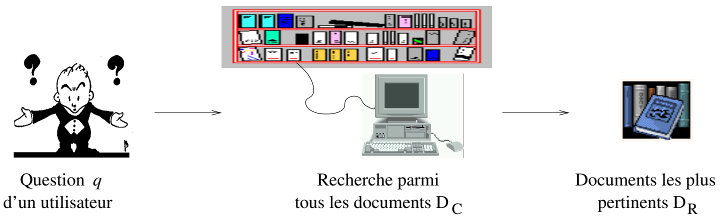


FIG. VI.1 – Un système de recherche avant l'enrichissement de la question

Cette entropie relative est évaluée par la mesure de la divergence de Kullback-Liebler (KLD) définie dans la théorie de l'information ([Cover et Thomas 1991]) :

$$KLD(D_R, D_C) = \sum_t \left\{ P_R(t) * \log \frac{P_R(t)}{P_C(t)} \right\} \tag{VI.3}$$

Les mots à considérer pour l'amélioration des requêtes sont ceux qui contribuent la plupart du temps à la divergence définie par cette distance. En pratique, il est fréquent que tous les termes de V n'apparaissent pas dans $R(q)$. C'est pourquoi, on note $v(R) \subset V$ le sous-ensemble du vocabulaire des termes qui apparaissent dans les documents représentés dans $R(q)$. Pour les termes n'appartenant pas à $v(R)$, on introduit une probabilité de back-off pour $P_R(t)$ sans quoi la mesure de divergence aurait une valeur infinie. Une possibilité pour calculer le back-off est de supposer qu'un terme non-occurrent a la même probabilité que dans la collection entière, comme dans ([Ponte et Croft 1998]). Une autre possibilité est, comme on le fait dans le modèle à base de mémoire cache en classification thématique, d'escompter les probabilités des termes observés tandis que la masse de probabilité récupérée de cette façon est redistribuée aux termes non-observés.

Cette approche donne le back-off suivant :

$$P_R(t) = \begin{cases} \mu \frac{tf}{NR} & si\ t \in v(R) \\ vP_C(t) & sinon \end{cases} \tag{VI.4}$$

où tf est le nombre d'occurrences du terme t dans les documents représentés dans $R(q)$, et NR est le nombre total de termes de l'ensemble des documents représentés dans $R(q)$, i.e. :

$$\sum_{t \in v(R)} tf = NR$$

Afin de s'assurer que les probabilités de tous les termes somment à 1, la contrainte suivante doit être respectée :

$$\mu + v \sum_{t \notin v(R)} P_C(t) = \mu + vA = 1$$

où :

$$A = \sum_{t \notin v(R)} P_C(t)$$

La divergence de Kullback-Liebler de l'equation VI.3 peut alors s'écrire :

$$KLD(D_R, D_C) = K_1 + K_2$$

$$K_1 = \sum_{t \in v(R)} \left\{ \mu \frac{tf}{NR} \log \frac{v \frac{tf}{NR}}{P_C(t)} \right\}$$

$$K_2 = \sum_{t \notin v(R)} \left\{ vP_C(t) \log \frac{vP_C(t)}{P_C(t)} \right\} = \sum_{t \notin v(R)} vP_C(t) \log v$$

avec :

$$v = \frac{1-\mu}{A}$$

donc on peut réécrire K_2 tel que :

$$K_2 = \sum_{t \notin v(R)} \frac{1-\mu}{A} P_C(t) \log \frac{1-\mu}{A}$$

En supposant que :

$$m(\mu) = \max_{t \notin v(R)} \frac{1-\mu}{A} P_C(t) \log \frac{1-\mu}{A}$$

les termes choisis pour l'amélioration de la requête sont ceux qui satisfont la condition :

$$\left\{ \mu \frac{tf}{NR} \log \frac{v \frac{tf}{NR}}{P_C(t)} \right\} > m(\mu)$$

Car μ n'influence pas l'ordre de $t \in v(R)$, les scores suivants peuvent être utilisés pour l'ordonnancement :

$$\sigma(t) = \mu \frac{tf}{NR} \log \frac{v \frac{tf}{NR}}{P_C(t)} \qquad \text{(VI.5)}$$

Les termes supérieurs sont choisis selon cet ordre pour l'expansion de requête. Les mêmes scores peuvent également être utilisés pour classer les termes choisis dans la requête augmentée. Dans ce cas le résultat dépend de la valeur de μ choisie.

Il convient de noter que même si $\mu = 1$, pour certains $t \in v(R)$ la valeur du $\sigma(t)$ pourrait être négative. Ceci se produit chaque fois que la probabilité de l'occurrence de t dans $v(R)$ est plus petite que celle correspondante dans la collection entière. Ceci implique que le log de la formule VI.5 peut être inférieur à 1.

4 Résultats obtenus sur TREC

L'objectif de la partie expérimentale est de vérifier l'hypothèse que les méthodes d'évaluation des termes, basées sur l'analyse des distributions, peuvent être utilisées pour montrer la pertinence de l'expansion automatique de requête sur la base initiale de Rocchio. Dans les expériences présentées, les scores produits par les fonctions distributionnelles ont été employés non seulement pour sélectionner mais aussi pour affecter un poids aux termes choisis. L'ensemble des expériences ont été réalisées avec les données des campagnes TREC-7 et TREC-8.

4.1 La campagne TREC

TREC (Text REtrieval Conference)[14] est une expérience à grande échelle faisant participer un certain nombre de groupes de recherche travaillant sur la recherche documentaire. TREC est co-sponsorisée par le "National Institute of Standards and Technology" (NIST)[15] et le "Defense Advanced Research Projects Agency" (DARPA)[16]. Il a débuté en 1992. Pour chaque TREC, le NIST fournit un ensemble de documents, de tests et de questions. Les participants utilisent leurs propres systèmes de recherche sur ces données, et renvoient au NIST une liste des documents obtenus avec leur système. Le NIST met les différents résultats en commun, juge les documents obtenus, et évalue les résultats. Le cycle de TREC se termine par un atelier qui est un forum où les participants partagent leurs expériences.

Les données de TREC-7 et TREC-8 ont été utilisées pour l'expérimentation. Ils comprennent le même ensemble de documents (i. e., disques 4 et 5 de TREC, contenant approximativement 2 Go de données) et différents ensembles de requêtes (respectivement les sujets 351-400 et 401-450). Les sujets de TREC-7 ont été décrits avec une moyenne de 57,6 termes, alors que la moyenne sur les sujets de TREC-8 était 51,8 termes. Le système de base utilisé dans toutes les expériences a été développé dans le contexte de la participation à TREC-8 de Messieurs Carpineto et Romano. Le système supprime les mots qui appartiennent à un anti-dictionnaire et extrait la racine des mots ("stemming"), en utilisant un lexique morphologique pour l'anglais ([Karp et al. 1992]). L'indexation de mot-clés a été exécutée pour toutes les données de test.

[14]http ://trec.nist.gov/
[15]http ://www.nist.gov/
[16]http ://www.darpa.mil/

4.2 Fonctions pour l'ordonnancement des documents

En plus de la méthode basée sur la distance de Kullback-Liebler, nous avons testé d'autres fonctions de classement des termes. La liste complète des fonctions testées dans l'expérience est la suivante (R indique le positionnement pseudo-approprié, C la collection entière, et $w(t)$ est le poids du terme t dans la collection) :

- Pondération de Rocchio :
$$score(t) = \sum_{k=1}^{r} w(t)_{Doc_k} \tag{VI.6}$$
- Valeur de sélection de Robertson (RSV) :
$$score(t) = \sum_{k=1}^{r} w(t)_{Doc_k} * P_R(t) \tag{VI.7}$$
- Mesure de CHI (CHI2) :
$$score(t) = \frac{[P_R(t) - P_C(t)]^2}{P_C(t)} \tag{VI.8}$$
- Variante de Doszkocs de la mesure de CHI (CHI1) :
$$score(t) = \frac{[P_R(t) - P_C(t)]}{P_C(t)} \tag{VI.9}$$
- Distance de Kullback-Liebler (KLD) :
$$score(t) = [P_R(t)] * \log\left[\frac{P_R(t)}{P_C(t)}\right] \tag{VI.10}$$

Nous avons considéré comme candidats à l'expansion tous les termes contenus dans R. Le nombre de documents pertinents utilisés à partir de la requête initiale pour construire R a été fixé à 10, alors que le nombre de termes de l'expansion considérés pour l'inclusion dans la requête augmentée est fixé à 40. L'évaluation des probabilités dans les formules ci-dessus est une question importante parce qu'elle pourrait affecter les performances des résultats. Pour estimer $P_X(t)$ dans le calcul de la distance KLD, comme mentionné dans la section 3, nous utilisons le quotient de la distribution des termes pour X, avec l'évaluation du maximum de vraisemblance de $P(t)$, i. e., le rapport entre la fréquence de t dans X, et le nombre de termes dans X. Pour estimer les probabilités dans les autres fonctions d'ordonnancement des termes nous avons également essayé différentes méthodes d'évaluation telles que le nombre de documents pseudo-appropriés qui contiennent le terme ([Buckley et Salton 1995]). Cette dernière méthode s'est avérée être assez mauvaise pour la recherche avec n'importe quelle fonction d'ordonnancement des termes, sauf avec la RSV. C'est pourquoi, nous avons choisi la probabilité basée sur les documents pseudo-appropriés seulement pour RSV ; en fait, c'est le choix recommandé pour la RSV dans [Robertson et al. 1995].

4.3 Système de classement des documents

Dans la première passe pour classer les documents les plus pertinents, le système utilise la formule suivante d'Okapi[17] ([Robertson et al. 1999]) pour calculer une mesure de similarité entre une requête q et un document d :

$$sim(q,d) = \sum_{t \in q \wedge d} w_{d,t} * w_{q,t} \tag{VI.11}$$

[17] Okapi est un système de recherche documentaire développé depuis 1982. Ce système a participé à toutes les campagnes TREC

avec :

$$w_{d,t} = \frac{(k_1+1) * f_{d,t}}{k_1 * \left[(1-b) + b * \frac{w_d}{moy\_w_d}\right] + f_{d,t}} \tag{VI.12}$$

et :

$$w_{q,t} = \frac{(k_3+1) * f_{q,t}}{k_3 + f_{q,t}} * \log \frac{N - f_t + 0,5}{f_t + 0,5} \tag{VI.13}$$

où k_1, k_3 et b sont des constantes respectivement aux valeurs $k_1 = 1,2$, $k_3 = 1000$ et $b = 0,75$. w_d est la taille du document exprimée en nombre de mots et moy_w_d est la taille moyenne des documents dans la collection entière. La valeur N est le nombre de documents dans la collection, f_t est le nombre de documents dans lesquels les termes t apparaissent, et $f_{x,t}$ est la fréquence du terme t dans le document d ou la requête q.

4.4 Evaluation des performances en utilisant Rocchio pour ordonner et affecter un poids aux termes

L'efficacité globale d'un système de recherche peut dépendre de nombreux facteurs. Le processus général de recherche utilisé pendant et après l'obtention des termes pour l'enrichissement de la question de l'utilisateur est présenté dans la figure suivante :

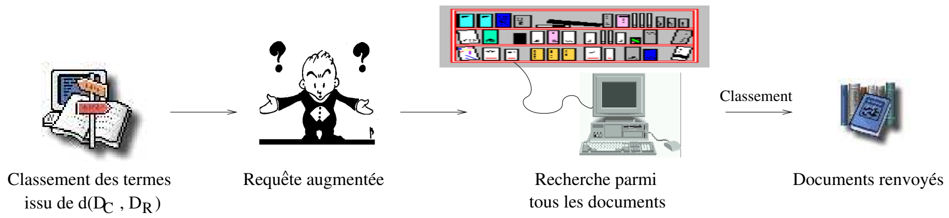


FIG. VI.2 – Un système de recherche pendant et après l'enrichissement de la question

Dans les expérimentations réalisées, nous avons uniquement fait varier la méthode employée pour choisir les termes de l'expansion de la requête, en maintenant constants les autres facteurs impliqués dans le processus de l'expansion de requête. Pour la ré-évaluation des poids des termes de la requête après leur sélection, nous avons uniformément utilisé la formule simplifiée de Rocchio (expression VI.1), avec $\alpha = 1$, $\beta = 1$. Les poids de la requête augmentée ont été alors employés dans l'expression VI.13 pour calculer la deuxième passe de recherche des documents.

Les quatre fonctions d'évaluation des termes des formules VI.7 à VI.10 peuvent être employées non seulement pour choisir les termes d'expansion mais aussi pour remplacer les poids dans la formule initiale de Rocchio VI.1. En d'autres termes, les poids du vecteur r dans la formule VI.1 peuvent être calculés en utilisant les scores des expressions VI.7 à VI.10, plutôt qu'en utilisant les poids indiqués par la formule VI.12, comme dans l'expression VI.6. Les poids du vecteur r et les poids de la requête initiale calculée par la formule VI.13 peuvent alors être normalisés, pour assurer l'uniformité.

En utilisant ce procédé pour déterminer les poids de la requête augmentée, dans l'expression VI.11, le rang de la deuxième passe peut être vu comme le produit de deux composantes, chacune impliquant un schéma de calcul des poids différent (c.-à-d., une évaluation des poids des termes concernant les documents et une évaluation des poids des termes de la requête, concernant la requête). L'utilisation d'un schéma de calcul des poids composé peut mieux refléter les différences d'importance des mêmes termes en ce qui concerne les documents dans la collection et en ce qui concerne la requête de l'utilisateur. La forme de notre fonction est conforme aux suggestions faites dans [Robertson 1990] et [Efthimiadis 1993].

Pour chaque requête, on a exécuté le système de classement complet pour chacune des cinq méthodes considérées. Les performances ont été mesurées avec les mesures standard d'évaluation de TREC qui sont présentées dans [Voorhees et Harman 1998] :
- le nombre de documents pertinents contenus dans les mille premiers renvoyés par le système (RET & REL),
- la précision moyenne (AV PREC),
- la précision à 11-points (11-PT-PREC),
- la R-précision (R-PREC),
- la précision à cinq documents renvoyés (PREC-AT-5),
- la précision à dix documents renvoyés (PREC-AT-10).

Dans les tables VI.1 et VI.2, nous présentons les performances des recherches de chaque méthode sur les collections de documents de TREC-7 et TREC-8. Nous rapportons également l'efficacité des recherches de la formule de base de Rocchio, et montrons l'amélioration des performances de chaque méthode d'ordonnancement avec la requête non-étendue, utilisée comme référence.

Les résultats montrent des taux d'amélioration significatifs. Une des raisons de cette amélioration est que si un terme candidat à l'expansion, pour une requête donnée, était correctement rangé, avant un autre terme, alors ce dernier devrait recevoir un poids proportionnellement plus élevé dans la requête augmentée. Tandis que si nous avions utilisé, pour la requête étendue, une fonction d'ordonnancement de poids qui calcule une valeur absolue de la qualité des termes, en ignorant l'information spécifique associée à la requête actuelle, comme on le fait dans la formule de base de Rocchio, alors le terme le meilleur pourrait recevoir un poids inférieur à celui qui est moins intéressant. Ainsi, l'utilisation d'un ordonnancement de termes distributionnel pour le re-calcul des poids de la requête, comme nous le faisons, peut gérer l'erreur d'assortiment possible entre la pertinence d'un terme à une requête donnée et la pertinence du même terme dans la collection.

	Requête initiale	Rocchio	RSV	CHI-2	CHI-1	KLD
RET&REL	2751	3009	3058	3007	3063	3202
		+9,38%	+11,16%	+9,31%	+11,34%	**+16,39%**
AV-PREC	0,2291	0,2625	0,2670	0,2555	0,2687	0,2873
		+14.54%	+16.53%	+11.50%	+17.26%	**+25.39%**
11-PT-PREC	0,2545	0,2806	0,2858	0,2747	0,2876	0,3054
		+10.25%	+12.28%	+7.93%	+12.99%	**+19.98%**
R-PREC	0,2711	0.2972	0.3077	0.2903	0.3012	0.3061
		+9.62%	+13.50%	+7.08%	+11.11%	**+12.93%**
PREC-AT-5	0.5480	0.5760	0.5680	0.5480	0.5560	0.6080
		+5.11%	+3.65%	0%	+1.46%	**+10.95%**
PREC-AT-10	0.5120	0.5240	0.5160	0.5080	0.5120	0.5240
		+2.34%	+0.78%	-0.78%	0.00%	**+2.34%**

TAB. VI.1 – Comparaison d'exécution de recherche sur TREC-7

	Requête initiale	Rocchio	RSV	CHI-2	CHI-1	KLD
RET&REL	2938	3111	2976	3173	3217	3269
		+5.89%	+1.29%	+8.00%	+9.50%	**+11.27%**
AV-PREC	0.2718	0.2972	0.2749	0.2798	0.2918	0.3053
		+9.33%	+1.14%	+2.94%	+7.35%	**+12.32%**
11-PT-PREC	0.2978	0.3174	0.2963	0.3007	0.3127	0.3229
		+6.57%	-0.52%	+0.96%	+4.99%	**+8.44%**
R-PREC	0.3168	0.3377	0.3215	0.3071	0.3369	0.3353
		+6.58%	+1.47%	-3.08%	+6.34%	+5.82%
PREC-AT-5	0.5960	0.6040	0.5800	0.5720	0.5480	0.6000
		+1.34%	-2.68%	-4.03%	-8.05%	+0.67%
PREC-AT-10	0.4920	0.5160	0.5320	0.4800	0.4840	0.5120
		+4.88%	**+8.13%**	-2.44%	-1.63%	+4.07%

TAB. VI.2 – Comparaison d'exécution de recherche sur TREC-8

Conclusion

Les méthodes présentées dans ce document s'appuient sur la modélisation statistique du langage, elles définissent notre contribution, tant d'un point de vue théorique que pratique, dans les domaines de la recherche documentaire et de la reconnaissance automatique de la parole. Les applications que nous avons développées en classification thématique, segmentation thématique et expansion de requête, ont, dans leur majorité, atteint les objectifs fixés.

La classification thématique, telle que nous l'avons perçue, a pour objet d'assigner un label thématique à un segment de texte parmi un ensemble prédéfini de labels possibles. Pour répondre à cet objectif, nous avons développé une méthode de reconnaissance thématique qui intègre un modèle cache avec un modèle de backing-off pour les n-grammes. Une mesure de divergence de Kullback-Leibler compare les distributions statistiques des mots au cours d'un document avec les distributions statistiques des mots dans un thème, ce qui permet une classification thématique dynamique. Cette méthode a été validée par des expérimentations sur des données du journal "Le Monde", des années 1987 à 1991, qui ont été réparties en 7 thèmes : Etranger, Histoire, Science, Sport, Economie, Culture et Politique. Comparée à des données de test labéllisées manuellement, cette méthode obtient un taux de classification thématique correcte de 74 %. De plus, ce modèle à base de mémoire cache a été combiné avec un ensemble d'unigrammes thématiques, par une règle de décision basée sur la logique floue. L'augmentation du taux de bonnes classifications, qui devient de 80 %, montre la complémentarité des deux modèles. Ce résultat satisfaisant nous permet de penser que le modèle cache est un outil fiable et performant. L'idée d'intégrer une mémoire cache ainsi qu'une sélection de mots clés offrent à ce modèle l'avantage de sélectionner les termes pertinents d'un document, d'un point de vue de la reconnaissance thématique. De plus, cela permet d'attribuer une étiquette thématique avec peu de données (5 mots clés). Les résultats sont d'autant plus probants que ces deux modèles sont simples à mettre en oeuvre, dans la mesure où l'on dispose de corpora d'apprentissage.

Le modèle à base de mémoire cache s'est également révélé efficace pour la classification thématique de textes issus de la retranscription de données dictées. Ce dernier point nous a permis d'envisager l'utilisation de cette classification dans le but d'améliorer les systèmes de reconnaissance automatique de la parole. L'objectif consiste à déterminer dynamiquement le thème d'un document en se servant des données du début de la reconnaissance, durant laquelle un modèle classique est utilisé. Nous avons alors montré la puissance prédictive des modèles bigrammes thématiques par le calcul de leur perplexité. Interpolé linéairement avec un modèle classique, cette approche permet un gain de perplexité de 8,7 % par rapport à un bigramme généraliste.

Dans ce mémoire, nous avons également abordé le problème de la segmentation thématique, que nous proposons de traiter en deux étapes : repérer un ensemble de ruptures thématiques candidates et les filtrer ensuite. Deux méthodes sont évaluées pour chacun de ces problèmes. Le modèle à base de mémoire cache, en premier lieu, est testé comme outil de repérage des ruptures candidates, pour sa capacité à discriminer les thèmes. Nous avons proposé une autre méthode basée sur des bigrammes à distance combinés avec une mémoire cache. Pour la sélection, nous utilisons la classification thématique du modèle cache, en considérant que deux segments adjacents qui reçoivent la même étiquette thématique sont de même thème, et donc la rupture proposée est annulée. Une autre solution de sélection repose sur une programmation dynamique où l'on recherche une séquence optimale de candidats en faisant varier la taille des historiques. A l'issue des expérimentations, nous avons constaté que la segmentation du modèle cache est la plus satisfaisante. Par le biais des diférents tests réalisés, nous avons montré qu'il est très délicat d'obtenir une segmentation thématique correcte sans connaissances préalables sur les thèmes. Le problème de la segmentation s'apparente ainsi à celui de la classification.

Le dernier problème que nous avons traité concerne l'expansion automatique de requête, par retour de pertinence. La méthode que nous proposons repose sur la divergence de Kullback-Leibler, issue de la théorie de l'information. Les termes de la requête étendue se déduisent des documents fournis par la requête initiale, en leur attribuant un score à partir de la distance entre les distributions statistiques des termes dans la collection entière et dans les documents renvoyés. Nous avons montré par une expérimentation sur TREC-7 et TREC-8 qu'elle peut être pertinente car elle est la seule méthode dont l'amélioration est régulière par rapport à la formule de Rocchio. Par ailleurs, cette méthode s'avère efficace non seulement pour la seule sélection des termes pertinents, mais aussi pour leur attribuer un score.

Dans le cadre d'une amélioration de la classification et de la segmentation thématique, les travaux en cours ont pour objet de donner une nouvelle représentation de la mémoire cache en l'enrichissant de nouveaux mots, appris sur un corpus. Ces mots auront été fréquemment rencontrés dans l'histoire de chacun des mots qui composent le cache. Dans cette optique, nous avons d'ores et déjà effectué les apprentissages relatifs à cette approche. A titre d'exemples, les 10 mots les plus fréquemment rencontrés dans un historique de 100 mots du mot "football" sont *club, match, Coupe, sport, football, stade, matches, clubs, finale et joueurs*, et ceux du mot "pic" sont *pic, taux, vallée, chercheurs, station, dollar, hausse, observatoire, atteindre et augmenté*. Ces historiques seront alors introduits dans la mémoire cache dès lors que le mot "football" ou "pic" sera rencontré dans le texte. Une autre piste qui nous semble importante concerne la sélection des mots clés. Cette étape est essentielle pour le modèle à base de mémoire cache, puisque la représentation des thèmes et du contexte du document repose sur eux. C'est en effet sur leur potentiel à discriminer les thèmes que s'appuie le modèle. Dans le travail que nous avons réalisé, les mots clés sont les mots les plus probables dans un thème, sauf s'ils appartiennent à une stop liste. Une autre perspective qui peut être donnée consiste en la pondération des mots du cache. Différents critères de pondération peuvent alors être envisagés (temporel, par classe syntaxique...). Les travaux futurs pourront également s'orienter vers la constitution de corpora thématiques plus importants, afin de disposer de suffisamment de données thématiques pour apprendre des modèles trigrammes. Dans ce cas, la classification thématique pourra être intégrée dans un système de reconnaissance de la parole, et le gain éventuel s'exprimera sous forme de taux de reconnaissance. Mais si l'on considère que le manque d'information sur les thèmes est un problème auquel il est difficile de répondre, une autre solution possible est d'intégrer les connaissances thématiques disponibles par une méthode plus adaptée que l'interpolation linéaire. Différentes possibilités pourront alors être testées.

Pour la poursuite du travail en expansion de requête, nous envisageons la possiblilité d'intégrer une méthode de classification qui constituerait une étape préalable à l'expansion. Le but serait la classification thématique des documents renvoyés par la requête initiale, afin de mieux cibler les sujets sous-jacents à la requête et donc mieux estimer les termes à y intégrer. Si l'on prend l'exemple du mot "avocat", il est possible de séparer les documents renvoyés en 2 classes (le fruit et le magistrat). On peut ensuite appliquer le processus d'expansion automatique de requête sur chaque classe, de sorte que chaque classe soit (équitablement ?) représentée dans la requête étendue. Il est fort probable que les termes choisis soient différents de ceux que l'on aurait obtenus en traitant les documents dans leur globalité. Plusieurs questions émergent : quelle est la meilleure classe ? Est-ce à l'utilisateur de choisir ? Est-ce qu'on obtient autant de nouvelles requêtes étendues que de classes ? Dans cette optique, la classification thématique des documents renvoyés serait une solution envisageable pour résoudre des problèmes liés aux requêtes mal exprimées ou lorsque le système n'a pas su correctement les prendre en compte.

Annexe A

ProbaSeg

ProbaSeg est le programme informatique, réalisé en langage C, qui prend en considération l'ensemble des problématiques abordées dans les chapitres IV et VI, concernant la classification thématique de segments de textes et la segmentation thématique de documents. Il prend en entrée un texte à traiter et effectue les analyses spécifiées dans un fichier "descripteur". Les tâches que ce programme peut effectuer sont les suivantes :

1. la classification thématique seule,
2. la segmentation thématique seule,
3. les deux.

Les paramètres nécessaires à l'entrée du programme sont les suivants :
- le fichier qui contient le vocabulaire,
- les fichiers des unigrammes thématiques,
- le type d'analyse souhaitée est "mot". En effet, les versions "lemmes" et "combinées" nécessitent un ensembles de pré-traitements coûteux en temps de calcul, et n'offrent pas une amélioration significative des résultats.
- le nombre de mots jusqu'auxquels le modèle cache ne sera pas pris en compte (cette valeur est généralement fixée à 5),
- le nombre de mots clés par thème (généralement fixé à 4000) et le nom de l'extension des fichiers,
- le fichier de la stop liste,
- la méthode de segmentation, si la segmentation est choisie.

Les résultats obtenus peuvent être édités dans un fichier par exemple pour les évaluations, ou visualisés graphiquement. Cette partie annexe du mémoire présente des copies d'écran de la version de démonstration de "ProbaSeg". Les trois premières figures montrent la fenêtre de spécification des options. Les figures 4 à 7 donnent les fenêtres de traitement de textes, écrits ou dictés, dans la tâche de classification thématique. Les dernières figures donnent une visualisation graphique des résultats de la segmentation thématique (signalée par des lignes pointillées verticalement).

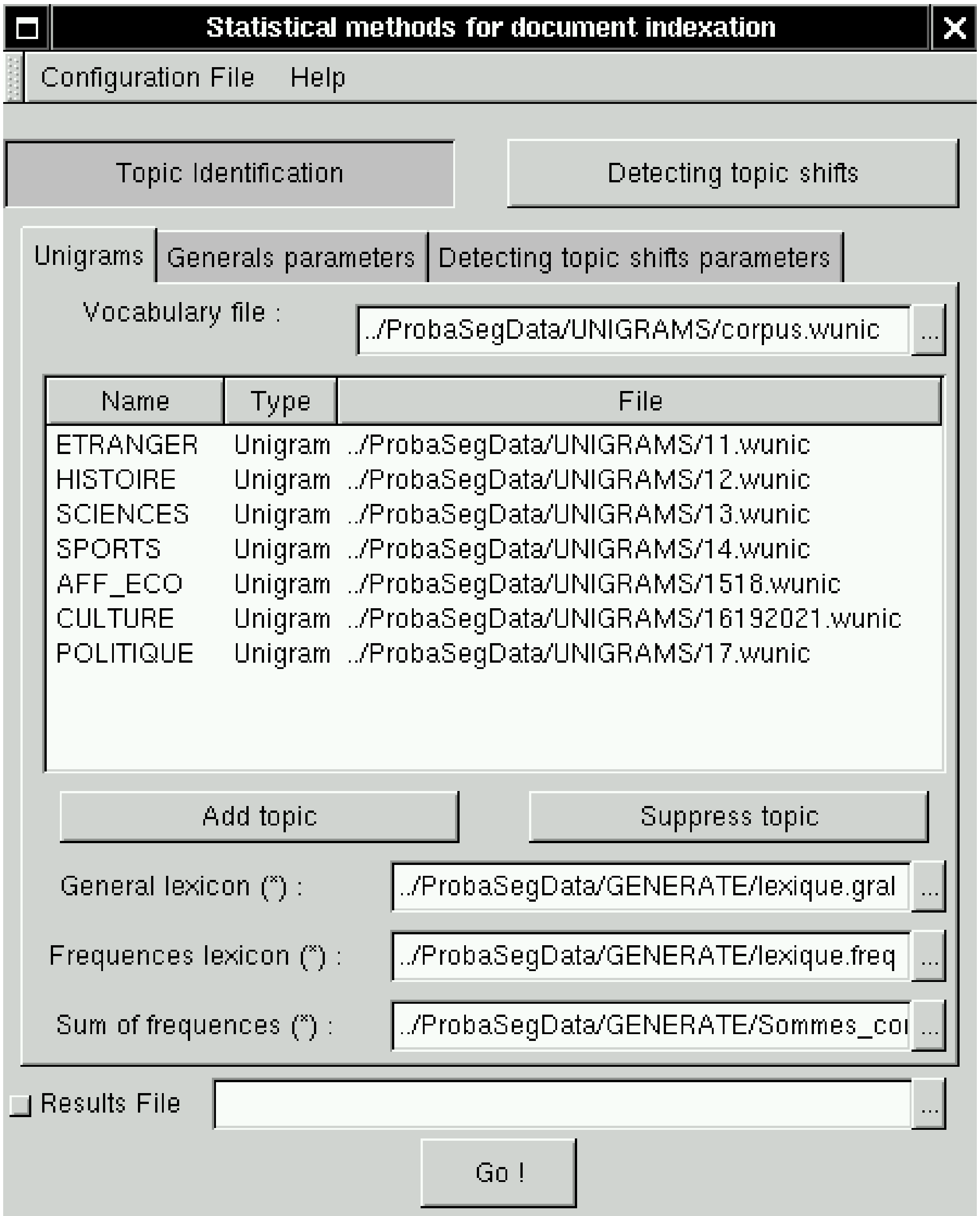


FIG. A.1 – Premier onglet de la fenêtre de spécification des options

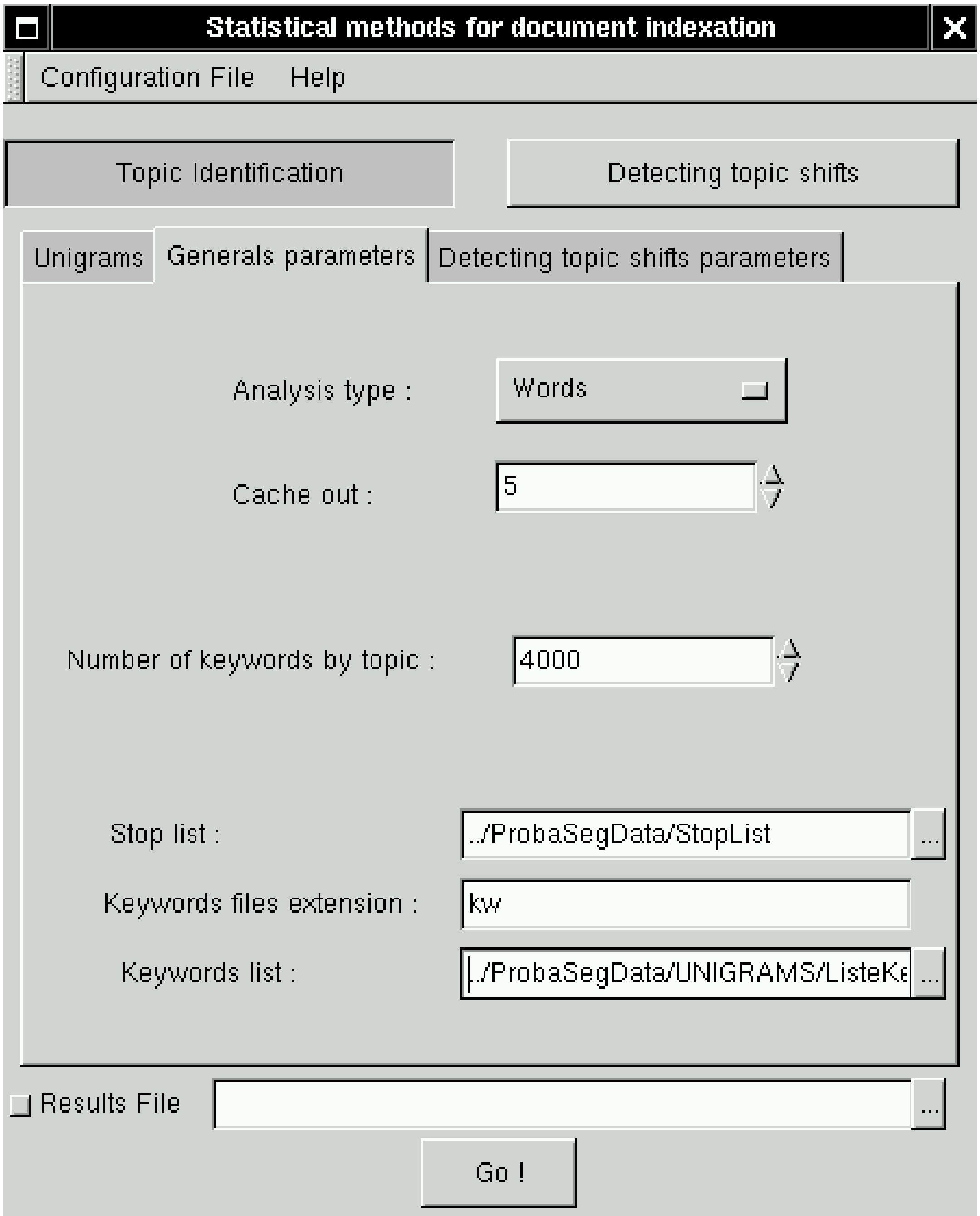


Fig. A.2 – Deuxième onglet de la fenêtre de spécification des options

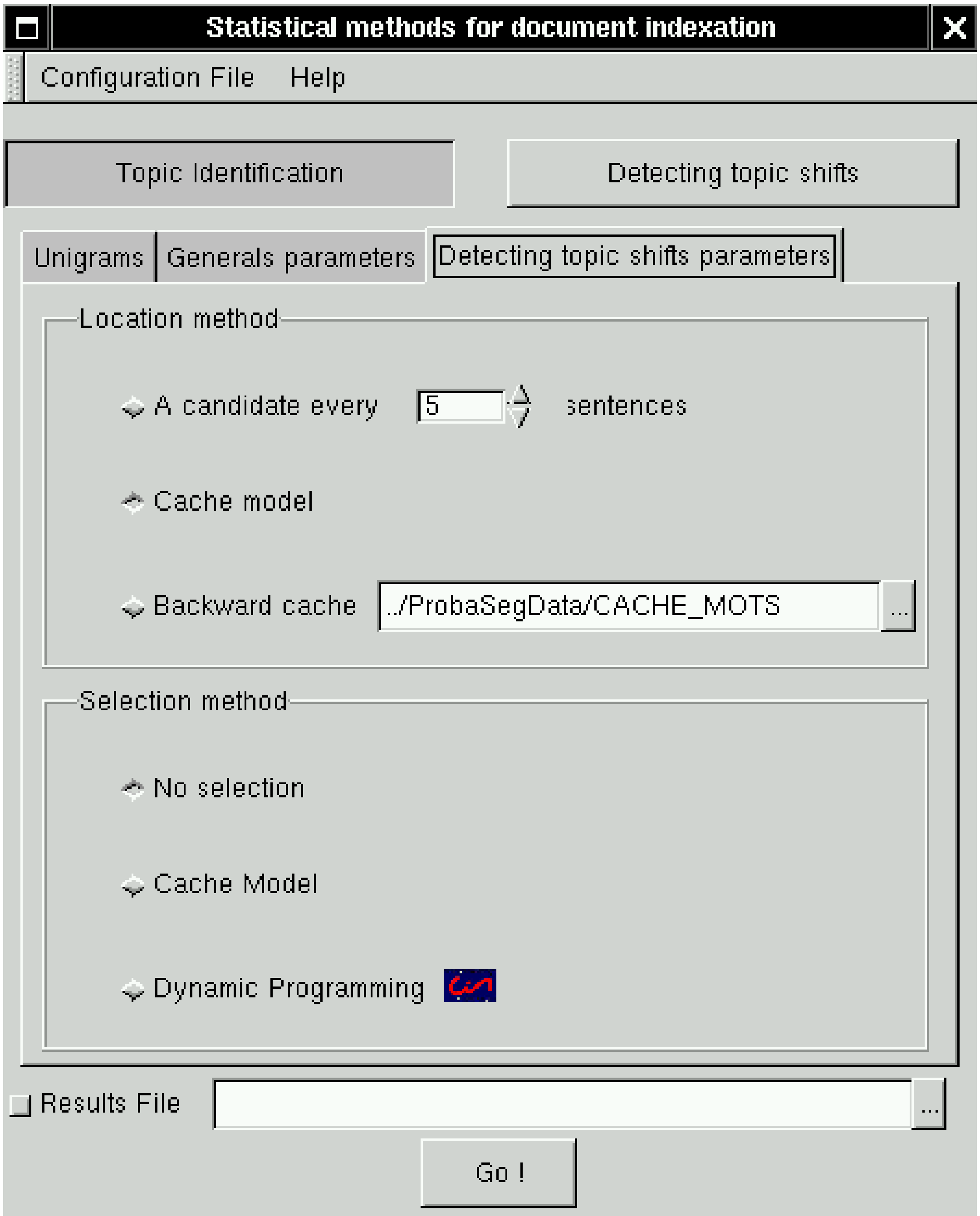


Fig. A.3 – Troisième onglet de la fenêtre de spécification des options

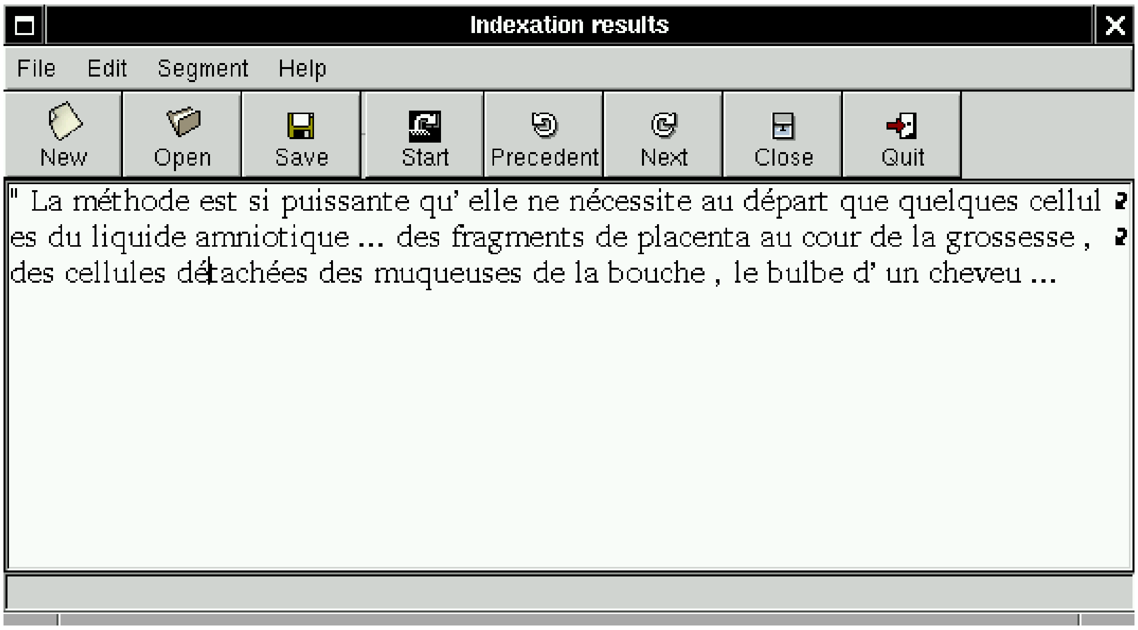


FIG. A.4 – Fênetre de traitement d'un texte issu du corpus du thème *Sciences*

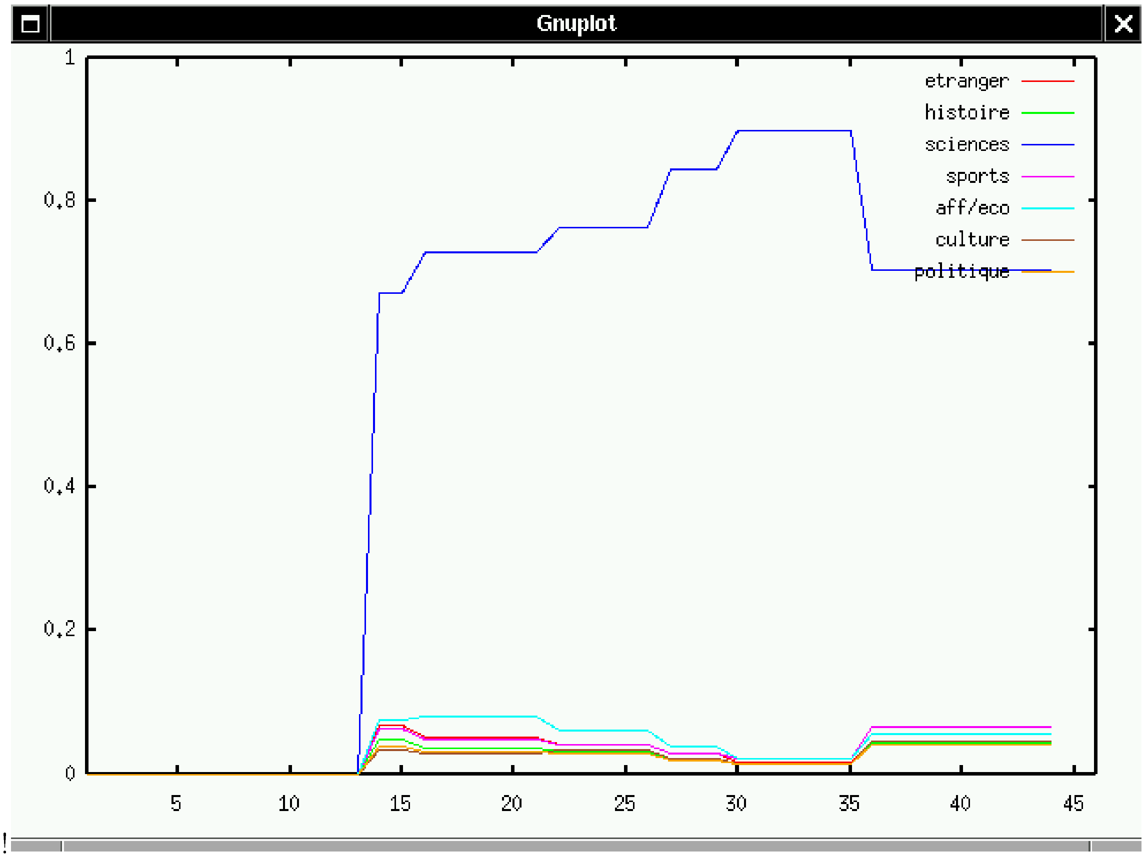


FIG. A.5 – Courbe de résultat obtenue sur le texte de la figure A.4

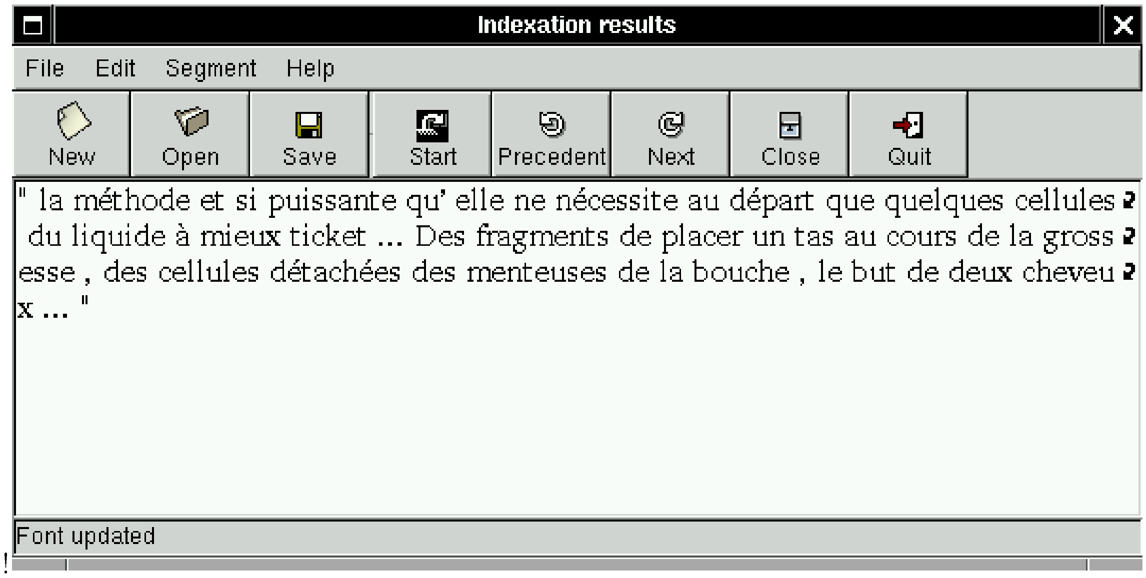


FIG. A.6 – Fênetre de traitement d'un texte issu du corpus du thème *Sciences* dicté au système *ViaVoice*

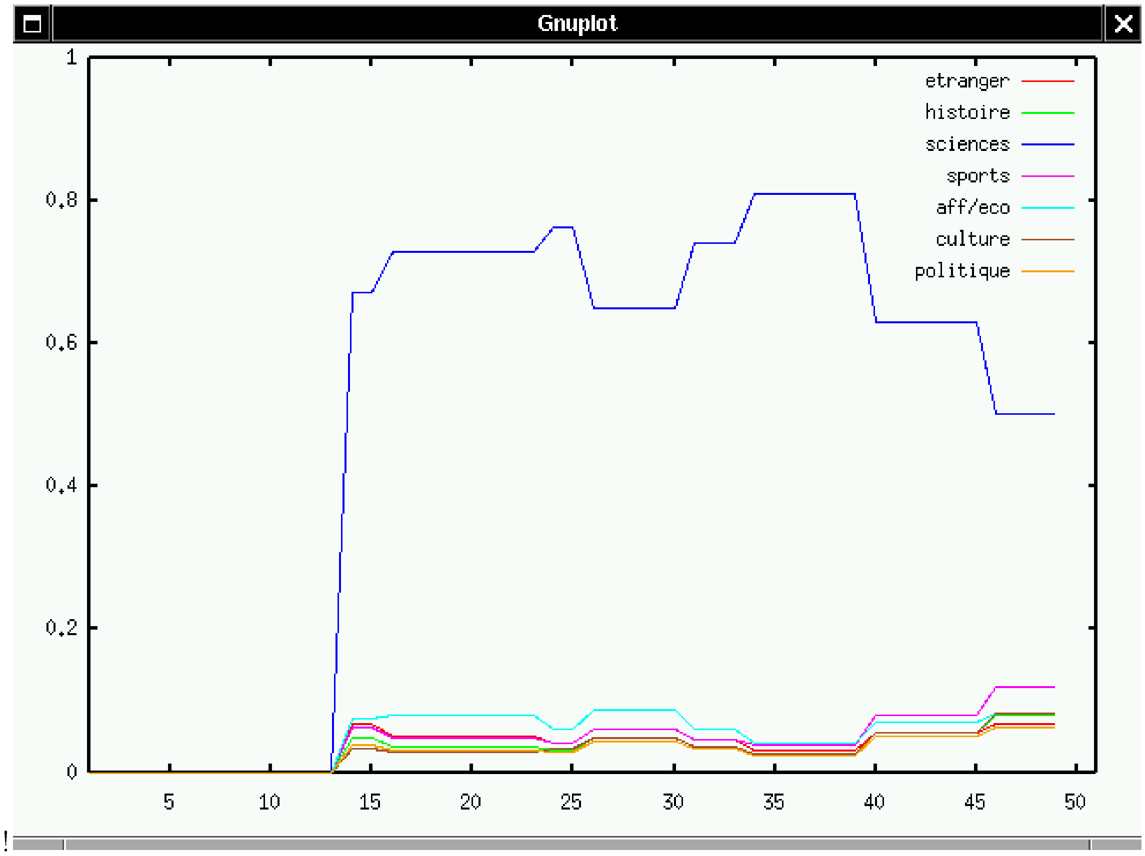


FIG. A.7 – Courbe de résultat obtenue sur le texte de la figure A.6

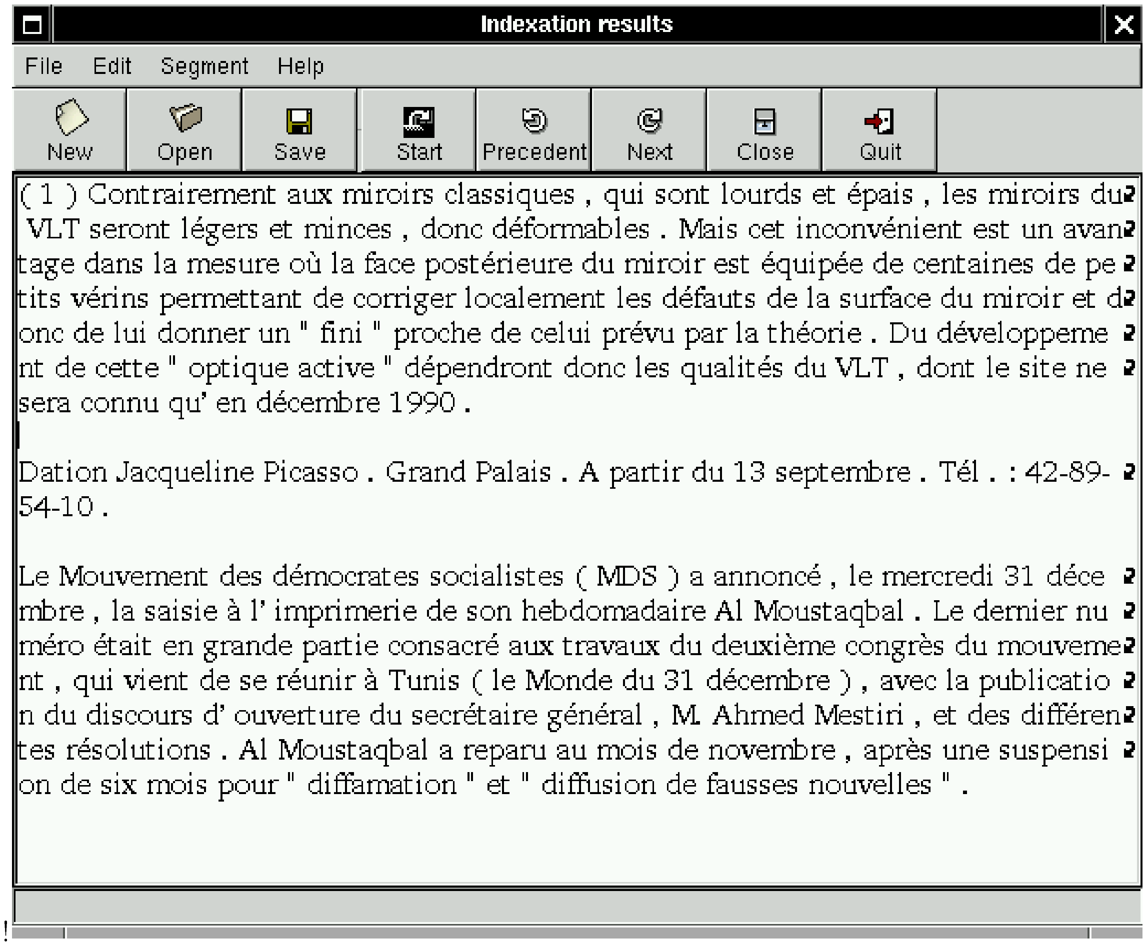


FIG. A.8 – Fenêtre de traitement d'un texte composé de 3 paragraphes issus des thèmes *Science*, *Culture* et *Etranger*.

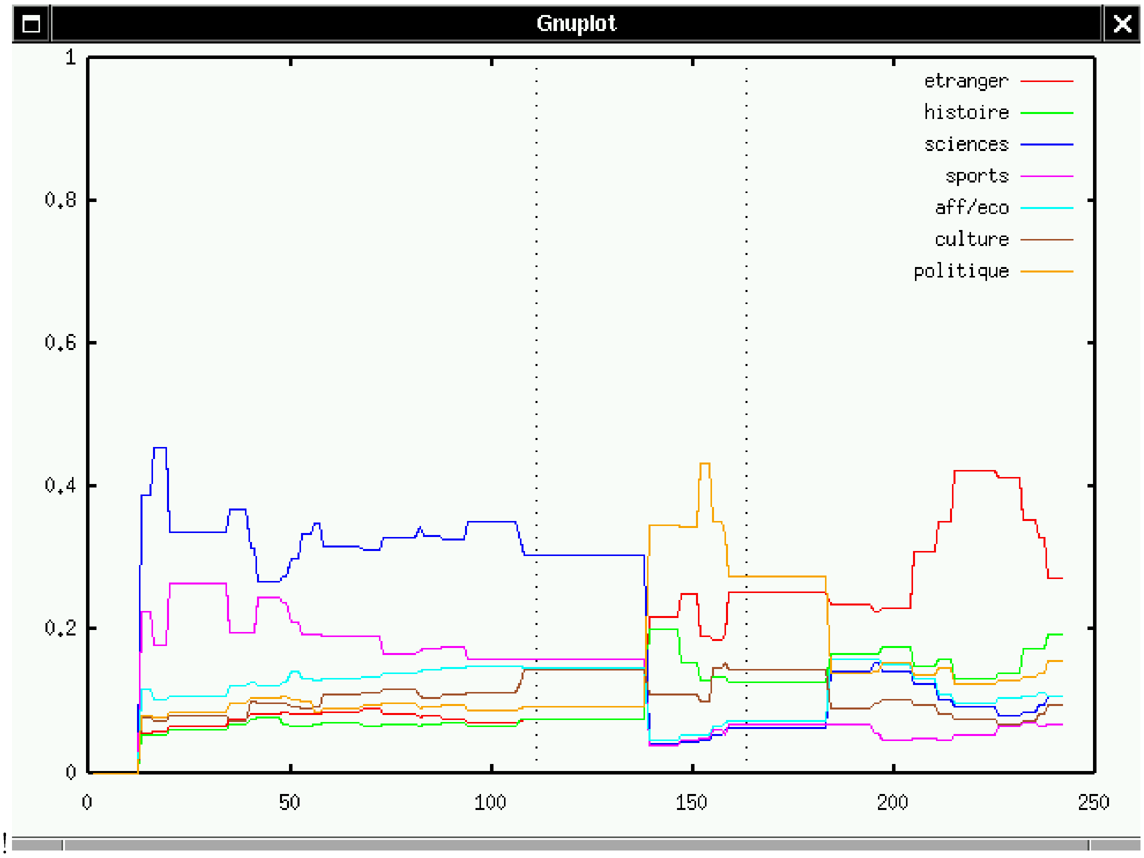


FIG. A.9 – Courbe de résultat obtenue sur le texte de la figure A.8. La première rupture indiquée est une fausse alarme. La seconde marque la rupture entre le premier et le deuxième paragraphe avec un retard de 2 phrases. La rupture entre le deuxième paragraphe et le troisième est omise.

Annexe B

Liste des étiquettes syntaxiques

Etiquette	Catégorie syntaxique
ADV	adverbe
ADVNE	ne
ADVPAS	pas
AFP	adjectif féminin pluriel
AFS	adjectif féminin singulier
AIND...	adjectif indéfini
AMP	adjectif masculin pluriel
AMS	adjectif masculin singulier
CHIF	chiffre ou nombre
COCO	conjonction de coordination
COSUB	conjonction de subordination
DET...	déterminant
DINT...	déterminant interrogatif
MOTINC	mot inconnu
NFP	nom féminin pluriel
NFS	nom féminin singulier
NMP	nom masculin pluriel
NMS	nom masculin singulier
PDEM...	pronom démonstratif
PIND...	pronom indéfini
PINT...	pronom interrogatif
PPER...	pronom personnel
PPOBJ...	pronom personnel objet
PREF...	pronom réfléchi
PREL...	pronom relatif
PREP	préposition
PREPADE	à de
PREPAU	au
PREPAUX	aux
PREPDES	des
PREPDU	du

Etiquette	Catégorie syntaxique
V...	verbe
VA...	auxiliaire avoir conjugué
VAINF	auxiliaire avoir à l'infinitif
VE...	auxiliaire être conjugué
VEINF	auxiliaire être à l'infinitif
VINF	verbe à l'infinitif
VPP...	participe passé
VPPRE	participe présent
XFAMIL	nom propre : nom de famille
XPAY...	nom propre : nom de pays
XPREF	prénom féminin
XPREM	prénom masculin
XSOC	nom propre : société
XVILLE	nom propre : localité
YPFAI	ponctuation faible , : ;()
YPFOR	ponctuation forte . ? !
ZTRM	début ou fin de phrase

TAB. B.1 – Etiquettes syntaxiques utilisées par le tagger du LIA, avec les "..." désignant les différentes sous-catégories (féminin pluriel, féminin singulier, masculin pluriel, masculin singulier pour les déterminants, adjectifs et pronoms et les différentes personnes pour les pronoms personnels et les verbes conjugués)

Publications personnelles

[Bigi 1997a] B. Bigi (1997). Combinaison de modèles de langages thématiques. Mémoire de DEA, Laboratoire d'Informatique d'Avignon.

[Bigi 1997b] B. Bigi (1997). Modèles de langage basés sur l'identification des thèmes. *Rencontre des Jeunes Chercheur en Parole*, La Rochelle.

[Bigi et al. 1997] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (1997). Combined models for topic spotting and topic-dependent language modeling. B. H. Huang Edited by S. Furui et N. Wu Chu, IEEE Signal Processing Society Publ, éditeur : *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 535–542.

[Bigi et al. 1998a] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (1998a). Combinaison de modèles de langage pour l'identification de thèmes. *XXIIèmes Journées d'Etudes sur la Parole*, pages 347–350, Martigny, Suisse.

[Bigi et al. 1998b] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (1998b). Detecting topic shifts using a cache memory. *5th International Conference on Spoken Language Processing, ICSLP-98*, Sydney, Australia.

[Bigi et al. 1999] B. Bigi (1999) Sélection dynamique de modèles de langage thématiques en RAP. *Rencontre des Jeunes Chercheur en Parole*, Avignon.

[Bigi et al. 2000a] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (2000a). A Fuzzy Decision Strategy for Topic Identification and Dynamic Selection of Language Models. Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal, 80(6).

[Bigi et al. 2000b] B. Bigi, R. De Mori et T. Spriet (2000b). Reconnaissance thématique à partir de textes dictés et Adaptation dynamique de modèles de langages thématiques. *XXIIIèmes Journées d'Etudes sur la Parole*, Aussois.

[De Mori et Bigi] R. De Mori et B. Bigi (A paraître). *Chapitre 5 : Principes de reconnaissances*. Ouvrage Traitement automatique du langage parlé.

Bibliographie

[Anastasakos et al. 1997] T. Anastasakos, J. Mc Donough et J. Makhoul (1997). Speaker adaptive training : a maximum likelihood approach to speaker normalisation. *IEEE Intermational Conference on Acoustics, Speech and Signal Processing*, pages 1043–1046, Munich, Germany.

[Apté et al. 1994] C. Apté, F. Damerau et S. Weiss (1994). Towards language independent automated learning of text categorization models. *7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 23–30, Dublin, Ireland.

[Bahl et al. 1993] L.R. Bahl, J. Bellegarda, P. de Souza, P. Gopalakrishnan, D. Nahamoo et M. Picheny (1993). Multonic Markov Word Models for Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing, 1(3) :334–344.

[Bahl et al. 1989] L.R. Bahl, P. Brown, P. Souza et R. Mercer (1989). A Tree-Based Statistical Language Model for Natural Language Speech Recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-37(7) :1001–1008.

[Bahl et al. 1983] L.R. Bahl, F. Jelinek et R. Mercer (1983). A Maximum Likelihood Approach To Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(2) :179–190.

[Bellot 2000] P. Bellot (2000). *Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire*. Thèse, Université d'Avignon et des Pays de Vaucluse, Avignon.

[Bellot et El-Bèze 2000] P. Bellot et M. El-Bèze (2000). *Classification et segmentation de textes par arbres de décision*. Techniques et Systèmes Informatiques, numèro spécial sur la Recherche d'Informations, Editions Hermès.

[Besling et Meier 1995] S. Besling et H. Meier (1995). Language Model Speaker Adaptation. *4th European Conference on Speech Communication and Technology*, volume 3, pages 1755–1758, Madrid, Spain.

[Bigi 1997] B. Bigi (1997). Combinaison de modèles de langages thématiques. Mémoire de DEA, Laboratoire d'Informatique d'Avignon.

[Bigi et al. 1997] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (1997). Combined models for topic spotting and topic-dependent language modeling. B. H. Huang S. Furui et N. Wu Chu, IEEE Signal Processing Society Publ, éditeur : *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 535–542.

[Bigi et al. 1998a] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (1998a). Combinaison de modèles de langage pour l'identification de thèmes. *XXIIèmes Journées d'Etudes sur la Parole*, pages 347–350, Martigny, Suisse.

[Bigi et al. 1998b] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (1998b). Detecting topic shifts using a cache memory. *5th International Conference on Spoken Language Processing, ICSLP-98*, Sydney, Australia.

[Bigi et al. 2000a] B. Bigi, R. De Mori, M. El-Bèze et T. Spriet (2000a). A Fuzzy Decision Strategy for Topic Identification and Dynamic Selection of Language Models. Special Issue on Fuzzy Logic in Signal Processing, Signal Processing Journal, 80(6).

[Bigi et al. 2000b] B. Bigi, R. De Mori et T. Spriet (2000b). Reconnaissance thématique à partir de textes dictés et Adaptation dynamique de modèles de langages thématiques. *XXIIIèmes Journées d'Etudes sur la Parole*, Aussois.

[Bimbot et al. 1997] F. Bimbot, M. El-Bèze et M. Jardino (1997). An alternative scheme for perplexity estimation. *ICASSP*, Munich, Germany.

[Brajnik et al. 1996] G. Brajnik, S. Mizzaro et C. Tasso (1996). Evaluating user interface to information retrieval systems : a case study on user support. *SIGIR'96*, pages 128–136, Zurich.

[Breiman et al. 1984] L. Breiman, J. Friedman, R. Olshen et C. Stone (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

[Brown et al. 1992] P.F. Brown, V. Della Pietra, J. L. P.V. de Souza et R. Mercer (1992). Class-Based n-gram Models of Natural Language. Computational Linguistics, 18(4) :467–479.

[Brugnara et Federico 1996] F. Brugnara et M. Federico (1996). Techniques for Approximating a Trigram Language Model. *International Conference of Spoken Language Processing*, Philadelphia, PA.

[Buckley et Salton 1995] C. Buckley et G. Salton (1995). Optimization of relevance feedback weights. *18th Annual International ACM SIGIR Conference on Research and Developpement in Information Retrieval,*, pages 351–357, Seattle.

[Carlson 1996] B.A. Carlson (1996). Unsupervided topic clustering of SWITCHBOARD speech messages. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 315–319, Atlanta GA.

[Carpineto et Romano 1998] C. Carpineto et G. Romano (1998). Effective reformulation of Boolean queries with concept lattices. *Third International Conference on Flexible Query-Answering Systems*, pages 83–94, Springer Verlag.

[Carpineto et Romano 1999] C. Carpineto et G. Romano (1999). Towards better techniques for automatic query expansion. *Third European Conference on Digital Libraries*, pages 126–141, Springer Verlag.

[Carpineto et al. 1998] C. Carpineto, G. Romano et R. De Mori (1998). Information term selection for automatic query expansion. *Seventh Text REtrieval Conference*, pages 363–370. NIST Special Publication.

[Chen et al. 1998] F. Chen, K. Seymore et R. Rosenfeld (1998). Topic adaptation for language modeling using unnormalized exponential models. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle WA.

[Chevalier et al. 1995] H. Chevalier, C. Ingold, C. Kunz, C. Moore, C. Roven, J. Yamron, B. Baker, P. Bamberg, S. Bridle, T. Bruce et A. Weader (1995). Large-Vocabulary Speech Recognition in Specialized Domains. *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 217–220, Detroit, MI.

[Church 1988] K.W. Church (1988). A stochastic parts program and noun phrase parser for unrestricted text. *Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas.

[Cooper et Byrd 1997] J. Cooper et R. Byrd (1997). Lexical navigation : visually prompted query expansion and refinement. *2nd ACM Digital Library Conference*, pages 237–246.

[Cover et Thomas 1991] T.M. Cover et J. Thomas (1991). *Elements of Information Theory*. Wiley.

[Dagan et al. 1994] I. Dagan, F. Pereira et L. Lee (1994). Similarity-based estimation of word co-occurence probabilities. *IEEE Conference of the Association for Computational Linguistics*, Socorro, New Mexico.

[Damnati 2000] G. Damnati (2000). *Modèles de langage et classification automatique pour la reconnaissance de la parole continue dans un contexte de dialogue oral homme-machine*. Thèse, Université d'Avignon et des Pays de Vaucluse, Avignon.

[De Loupy et al. 1999] C. De Loupy, P. Bellot, M. El-Bèze et P. Marteau (1999). Query Expansion and Classification of Retrieved Documents. *Seventh Text Retrieval Conference*, pages 443–450, Gaithersburg MD (USA).

[De Mori 1998] R. De Mori (1998). *SPOKEN DIALOGUES WITH COMPUTERS*. Academic Press.

[De Mori et Bigi A paraître] R. De Mori et B. Bigi (A paraître). *Chapitre 5 : Principes de reconnaissances*. Ouvrage Traitement automatique du langage parlé, Editions Hermès.

[Della Pietra et Della Pietra 1994] S.A. Della Pietra et V. Della Pietra (1994). Statistical modeling using maximum entropy. Research Report, IBM.

[Della Pietra et al. 1992] S.A. Della Pietra, V. Della Pietra, R. Mercer et S. Roukos (1992). Adaptive Language Model Estimation using Minimum Discrimination Estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 633–636, San Francisco, CA.

[Doszcocks 1978] T.E. Doszcocks (1978). AID : an associative interactive dictionary for online searching. Online Review, 2(2) :163–174.

[Dugast et al. 1995] C. Dugast, X. Aubert et R. Kneser (1995). The Philips Large-Vocabulary Recognition System for American English, French and German. *Eurospeech*, Madrid, Espagne.

[Efthimiadis 1993] E. Efthimiadis (1993). A User-centered evaluation of ranking algorithms for interactive query expansion. *16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–159, Pittsburg, PA.

[El-Bèze et al. 1997] M. El-Bèze, M. Jardino et F. Bimbot (1997). Une approche alternative pour le calcul de la perplexité. *Actes des premières JST Francil*, pages 79–84, Avignon.

[Federico 1996] M. Federico (1996). Bayesian Estimation Methods for N-gram Language Model Adaptation. *International Conference of Spoken Language Processing*, Philadelphia, PA.

[Federico et al. 1995] M. Federico, M. Cettolo, F. Brugnara et G. Antoniol (1995). Language Modeling for Efficient Beam-Search. Computer Speech and Language, 9 :353–379.

[Ferret 1998] O. Ferret (1998). How to thematically segment texts by using lexical cohesion. *ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 1481–1483, Melbourne, Australie.

[Ferret et Grau 1998] O. FERRET et B. GRAU (1998). A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts. *Actes European Conference on Artificial Intelligence*, pages 155–159, Brighton, Grande-Bretagne.

[Fohr et al. 1997] D. FOHR, J. HATON, J. MARI, K. SMAÏLI et I. ZITOUNI (1997). MAUD : Un prototype de machine à dicter vocale. *Actes des premières JST Francil*, pages 25–30, Avignon.

[Foote et al. 1997] J.T. FOOTE, S. YOUNG, G. JONES et K. S. JONES (1997). Unconstrained keyword spotting using phone lattices with applications to spoken document retrieval. Computer Speech and Language, 11(3) :207–224.

[Gauvain et Lee 1994] J.L. GAUVAIN et C. LEE (1994). Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, 2 :291–298.

[Giachin 1995] E. GIACHIN (1995). Phrase Bigrams For Continuous Speech Recognition. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, pages 225–229, Detroit, MI.

[Greiff et Croft 1999] W.R. GREIFF et W. CROFT (1999). PIC Matrices : A computationally Tractable Class of Probabilistic Query Operators. ACM Transactions on Information Systems, 17(4) :367–405.

[Harman 1992] D. HARMAN (1992). Relevance feedback and other query modification techniques. W.B. FRAKES & R. BAEZA-YATES, éditeur : *Information Retrieval - Data Structures and Algorithms*, pages 241–263, Englewood Cliffs :Prentice Hall.

[Hawking et al. 1998] D. HAWKING, P. THISTLEWAITE et N. CRASWELL (1998). ANU/ACSys TREC-6 Experiments. HARMAN EDITOR D.K. HARMAN, éditeur : *Sixth Text REtrieval Conference*.

[Heinonen 1998] O. HEINONEN (1998). Optimal multi-paragraph text segmentation by dynamic programming. *36th Annual Meeting of the Association for Computational Linguistics*, pages 1484–1486, Montréal, Québec.

[Ikehare et al. 1996] S. IKEHARE, S. SHIRAI et H. UCHINO (1996). A statistical method for extracting uninterrupted collocations from very large corpora. *COLING*, pages 574–579.

[Imai et al. 1997] T. IMAI, R. SCHWARTZ, F. KUBALA et L. NGUYEN (1997). Improved Topic Discrimination of Broadcast News Using a Model of Multiple Simultaneous Topics. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 727–730, Munich, Germany.

[Isotani et Matsunaga 1994] ISOTANI et S. MATSUNAGA (1994). A stochastic language model for speech recognition integrating local and global constraints. *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 5–8, Adelaide, Australia.

[Iyer et al. 1994] R. IYER, M. OSTENDORF et J. ROHLICEK (1994). Language Modeling with Sentence-Level Mixtures. *ARPA Human Language Technology Workshop*, pages 82–86, Plainsboro, NJ.

[Jardino 1998] M. JARDINO (1998). Evaluation de modèles de langage à base de trigrammes de classes et de mots, avec le Jeu de Shannon. *XXIIèmes Journées d'Etudes sur la Parole*, pages 363–366, Martigny, Suisse.

[Jelinek 1976] F.J. JELINEK (1976). Continuous Speech Recognition by Statistical Methods. Institute of Electrical and Electronic Engineers, 4 :532–556.

[Jelinek 1990] F.J. Jelinek (1990). Self-Organized Language Modeling for Speech Recognition. Alex Weibel et K. Lee, éditeur : *Readings in Speech Recognition*, pages 450–505. Morgan Kaufmann, Los Altos, CA.

[Jelinek 1991] F.J. Jelinek (1991). Up from trigrams ! The struggle for improved language models. *European Conference on Speech Communication and Technology*, pages 1037–1040, Genova, Italy.

[Jelinek 1997] F.J. Jelinek (1997). *STATISTICAL METHODS FOR SPEECH RECOGNITION*. The MIT Press.

[Jelinek et Lafferty 1991] F.J. Jelinek et J. Lafferty (1991). Computation of the probability of initial substring generation by stochastic context free grammars. Computational linguistics, 17(3) :315–323.

[Jelinek et Mercer 1980] F.J. Jelinek et R. Mercer (1980). Interpolated estimation of Markov source parameters from sparse data. *Pattern Recognition in Practice*, pages 381–397, Amsterdam, Holland.

[Jobbins et Evett 1998] A.C. Jobbins et L. Evett (1998). Text segmentation using reiteration and collocation. pages 614–618.

[Junker et Abecker 1997] M. Junker et A. Abecker (1997). Exploiting thesaurus knowledge in rule induction for text classification. *Recent Advances in Natural Language Processing*, pages 202–207, Tzigov Chark, Bulgaria.

[Junqua et Haton 1996] J.C. Junqua et J. Haton (1996). *ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION*. Kluwer Academic.

[Karp et al. 1992] D. Karp, Y. Schabes, M. Zaidel et D. Egedi (1992). A freely available wide coverage morphological analyzer for English. *14th International Conference on Computational Linguistics*, Nantes.

[Katz 1987] S.M. Katz (1987). Estimation of probabilities from sparse data for the language model component of speech recognizer. IEEE Trans. Acoust. Speech and Signal Proc., ASSP-35(3) :400–410.

[Kenne et al. 1995] P.E. Kenne, M. O'Kane et H. Pearcy (1995). Language Modeling of Spontaneous Speech in a Court Context. *4th European Conference on Speech Communication and Technology*, volume 3, pages 1801–1804, Madrid, Spain.

[Khudanpur et Wu 1999] S.P. Khudanpur et J. Wu (1999). A Maximum Entropy Language Model Integrating N-Gram and Topic Dependencies for Conversational Speech Recognition. *IEEE International Conference On Acoustics, Speech, And Signal Processing*, volume 1, pages Paper number 2192, Phoenix, Arizona.

[Kneser 1996] R. Kneser (1996). Statistical language modeling using a variable context. *IEEE International Conference on Spoken Language Processing*, pages 494–497, Philadelphia, Penssylvania, USA.

[Kneser et Peters 1997] R. Kneser et J. Peters (1997). Semantic clustering for adaptive language modeling. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 779–783, Munich, Germany.

[Kneser et Steinbiss 1993] R. Kneser et V. Steinbiss (1993). On the Dynamic Adaptation of Stochastic Language Models. *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 586–588, Minneapolis, MN.

[Kowalski 1997] G. Kowalski (1997). *Information retrieval systems - Theory and implementation*. Kluwer Academic Publishers, ISBN-0-7923-9926-9.

[Krovetz 1995] R. Krovetz (1995). *Word sense disambiguation for large text database*. Thèse, University of Massachussetts, Amhest, MA, USA.

[Kuhn et De Mori 1990] R. Kuhn et R. De Mori (1990). A Cache-Based Natural Language Model for Speech Recognition. IEEE Trans. Pattern Anal. Machine Intell., PAMI-12(6) :570–582.

[Lau et al. 1993] R. Lau, R. Rosenfeld et S. Roukos (1993). Trigger-based Language Models : A maximum Entropy Approach. *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 45–48, Minneapolis, MN.

[Lee et al. 1996] C.H. Lee, F. Soong et K. Paliwal (1996). *AUTOMATIC SPEECH AND SPEAKER RECOGNITION : ADVANCED TOPICS*. Kluewer.

[Leggeter et Woodland 1995] C.J. Leggeter et P. Woodland (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 9 :171–185.

[Lewis et Gale 1994] D.D. Lewis et W. Gale (1994). A sequential algorithm for training text classifiers. *7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, Ireland.

[Li et Yamamishi 1997] H. Li et K. Yamamishi (1997). Documentation classification using a finite mixture model. *Conference of the Association for Computational Linguistics*, pages 39–47, Madrid, Spain.

[Mahajan et al. 1999] M. Mahajan, D. Beeferman et X. Huang (1999). Improved Topic-Dependent Language Modeling using Information Retrieval Techniques. *IEEE International Conference On Acoustics, Speech, And Signal Processing*, volume 1, pages Paper number 2391, Phoenix, Arizona.

[Mari et Haton 1994] J.F. Mari et J. Haton (1994). Automatic word recognition based on second-order hidden Markov models. *International Conference of Spoken Language Processing*, pages 274–277, Yokohama, Japan.

[McCandless et Glass 1994] M.K. McCandless et J. Glass (1994). Empirical Acquisition of Language Models for Speech Recognition. *International Conference of Spoken Language Processing*, volume 2, pages 835–838, Yokohama, Japan.

[McDonough et Gish 1994] J. McDonough et H. Gish (1994). Issues in topic identification on the switchboard corpus. *International Conference on Spoken Language Processing*, pages 2163–2166, Yokohama, Japan.

[Merialdo 1994] B. Merialdo (1994). Tagging English text with a probabilistic moddel. Computational Linguistics, 20(2) :155–172.

[Miller et al. 1999a] D.R.H. Miller, T. Leek et R. Schwartz (1999a). BBN at TREC7 : Using Hidden Markov Models for Information Retrieval. *Text REtrieval Conference TREC-7, NIST special publication 500-242*, pages 133–142, Gaithersburg, USA.

[Miller et al. 1999b] D.R.H. Miller, T. Leek et R. Schwartz (1999b). A Hidden Markov Model Information Retrieval System. *ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221, Berkeley, USA.

[Murveit et al. 1994] H. Murveit, P. Monaco, V. Digalakis et J. Butzberger (1994). Techniques to Achieve an Accurate Real-Time Large-Vocabulary Speech Recognition System. *ARPA Human Language Technology Workshop*, pages 368–373, Plainsboro, NJ.

[Ney et al. 1994] H. Ney, U. Essen et R. Kneser (1994). On structuring probabilistic dependences in stochastic language modelling. Computer Speech and Language, 8 :1–38.

[Ohtsuki et al. 1998] K. Ohtsuki, T. Matsutoka, S. Matsunaga et S. Furui (1998). Topic Extraction with multiple topic-words in broadcast news. *IEEE International Conference On Acoustics, Speech, And Signal Processing*, Seattle, WA.

[Peirera et al. 1993] F.C.N Peirera, N. Tishby et Lee (1993). Distributional clustering of English words. *Conference of the Association for Computational Linguistics*, pages 183–190, Columbus, Ohio, USA.

[Peskin et al. 1996] Peskin, S. Conolly, L. Gillick, S. Lowe, D. McAllaster, V. van Mulbregt et S. Wegmann (1996). Improvements in SWITCHBOARD recognition and topic identification. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 303–306, Atlanta GA.

[Peters et Kneser 1997] J. Peters et R. Kneser (1997). Semantic clustering for adaptative language modeling. *IEEE International Conference On Acoustics, Speech, And Signal Processing*, pages 779–782.

[Ponte et Croft 1998] J. Ponte et B. Croft (1998). A language modeling approach to information retrieval. *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281.

[Robertson 1990] S.E. Robertson (1990). On term selection for query expansion. Journal of Documentation, 46(4) :359–364.

[Robertson et al. 1999] S.E. Robertson, S.Walker et M. Beaulieu (1999). Okapi at TREC-7 : Automatic ad hoc, filtering, VLC, and interactive track. Seventh Text REtrieval Conference, 500(242) :253–264.

[Robertson et al. 1995] S.E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu et M. Gatford (1995). Okapi at TREC-3. *Third Text REtrieval Conference*, pages 109–126.

[Rocchio 1971] J.J. Rocchio (1971). Relevance Feedback in Information Retrieval. G. Salton, éditeur : *The SMART Retrieval System*, pages 313–323, Englewood Cliffs, N.J. :Prentice-Hall, Inc.

[Rosenfeld 1994] R. Rosenfeld (1994). *Adaptive Statistical Language Modeling : A Maximum Entropy Approach*. Thèse, School of Computer Science - Carnegie Mellon University, Pittsburgh, PA 15213.

[Rosenfeld 1995] R. Rosenfeld (1995). The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation. *ARPA Spoken Language Technology Workshop*, pages 47–50, Austin, Texas.

[Rosenfeld et Huang 1992] R. Rosenfeld et X. Huang (1992). Improvements in stochastic language modeling. *DARPA Speech and Natural Language Workshop*, San Mateo, California. Morgan Kaufman.

[Roses et al. 1991] R.C. Roses, E. Chang et R. Lippmann (1991). Techniques for information retrieval from Voice messages. *International conference on acoustics, speech and signal processing*, pages 317–320, Toronto, Canada.

[Salton et Buckley 1988] G. Salton et C. Buckley (1988). Term-weighting approaches in automatic text retrieval. Proceedings & Management, 24 :513–523.

[Salton et Buckley 1990] G. Salton et C. Buckley (1990). Improving retrieval Performance by Relevance Feedback. Journal of the American Society for Information Sciences, 41 :288–297.

[Salton et McGill 1983] G. Salton et M. McGill (1983). *Introduction to modern information retrieval*. Mc Gow Hill, New York, USA.

[Salton et al. 1975] G. Salton, A. Wong et C. Yang (1975). A vector space model for automatic indexing. Communications of the ACM, 18 :613–620.

[Sarukkai et Ballard 1997] R.R. Sarukkai et D. Ballard (1997). Word set probability boosting for improved spontaneous dialog recognition. IEEE Transactions on Speech and Audio Processing, SAP-5(5) :438–445.

[Schukat-Talamazzini et al. 1994] E.G. Schukat-Talamazzini, T. Kuhn et H. Niemann (1994). Speech recognition for spoken dialog systems. R. De Mori H. Niemann et G. Hahnrieder, éditeur : *Progress and Prospects of Speech Research and Technology*, Sainks Augustin, Germany. Infix.

[Seymore et al. 1998] K. Seymore, S. Chen et R. Rosenfeld (1998). Nonlinear interpolation of topic models for language model adaptation. *IEEE International Conference on Spoken Language Processing*, Sydney AUS.

[Seymore et Rosenfeld 1996] K. Seymore et R. Rosenfeld (1996). Scalable Backoff Language Models. *International Conference of Spoken Language Processing*, Philadelphia, PA.

[Shannon 1948] C.E. Shannon (1948). A mathematical theory of communication. Bell Systems Technical Journal, 27 :379–423.

[Shannon 1951] C.E. Shannon (1951). Prediction and entropy of printed English. Bell Systems Technical Journal, pages 20–64.

[Spriet et El-Bèze 1998] T. Spriet et M. El-Bèze (1998). Introduction of rules into a stochastic approach for language modeling. Springer Verlag K.M. Ponting, éditeur : *Computational Models of Speech Pattern Processing*, volume NATO ASI series F, Berlin New York.

[Steinbiss et Kneser 1997] V. Steinbiss et R. Kneser (1997). On the dynamic adaptation of stochastic language modeling. *IEEE International Conference On Acoustics, Speech, And Signal Processing*, volume II, pages 586–589.

[Stuart et al. 1990] L.C. Stuart, M. Fung, L. Appelbaum et R. Tong (1990). Classification trees for information retrieval. *8th. International Workshop on machine learning*, Evanston, Illinois.

[Tillman et Ney 1996] C. Tillman et H. Ney (1996). Selection criteria for word trigger pairs in language modeling. L. Miclet et C. D. la Higuera, éditeur : *Grammatical Inference : Learning Syntax from Sentences, Third International Colloquim*, pages 98–106, Montpellier.

[Turtle et Croft 1990] H. Turtle et W. Croft (1990). Inference networks for document retrieval. J.L. Vidick, éditeur : *Thirteenth International Conference on Research and Development in Information Retrieval*, pages 1–24, New York.

[van Mulbregt et al. 1998] P. van Mulbregt, I. Carp, L. Gillick, S. Lowe et J. Yamron (1998). Text Segmentation and Topic tracking on Broadcast news via Hidden Markov Model approach. *5th International Conference on Spoken Language Processing*, pages Paper number 116, Sydney, Australia.

[Van Rijsbergen 1979] C. Van Rijsbergen (1979). *Information Retrieval*. Buttherwords, London.

[Van Rijsbergen 1986] C. Van Rijsbergen (1986). A non-classical logic for information retrieval. Computer journal, 29 :481–485.

[Voorhees et Harman 1998] E. Voorhees et D. Harman (1998). Overview of the Seventh Text REtrieval Conference. *Seventh Text REtrieval Conference*, pages 1–24. NIST Special Publication.

[Witten et Bell 1991] I.H. Witten et T. Bell (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text comparison. IEEE Transactions Information Theory, 34(4) :1085–1094.

[Wong et Yao 1995] S.K.M. Wong et Y. Yao (1995). On modeling information retrieval with probabilistic inference. ACM Transactions on Information Systems, 13(1) :38–63.

[Wright et al. 1995] J.H. Wright, M. Carey et E. Parris (1995). Improved topic spotting through statistical modelling of keyword dependencies. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages Paper number 313, Detroit, MI.

[Wright et al. 1996] J.H. Wright, M. Carry et E. Parris (1996). Statistical models for topic identification using phoneme substring. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 307–310, Atlanta GA.

[Yaari 1997] Y. Yaari (1997). Segmentation of expository texts by hierarchical agglomerative clustering. *Recent Advances in Natural Language Processing*, pages 59–65, Tzigov Chark, Bulgarie.

[Yamashita et al. 1998] Y. Yamashita, T. Tsunekawa et R. Mizoguchi (1998). Topic recognition for news speech based on keyword spotting. *5th International Conference on Spoken Language Processing*, pages Paper number 23, Sydney, Australia.

[Yamron et al. 1997] J.P. Yamron, I. Carp, L. Gillick, S. Lowe et P. van Mulbregt (1997). Event tracking and Text Segmentation via Hidden Markov Models. *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 519–526, Santa Barbara, CA. S. Furui, B.H. Huang and Wu Chu, IEEE Signal Processing Society Publ, NY.

[Yang et al. 1999] K. Yang, K. Maglaughlin, L. Meho et R. Summer Jr (1999). IRIS at TREC-7. E.M. Voorhees et D. Harman, éditeur : *Seventh Text REtrieval Conference*, pages 555–566. NIST Special Publication.

[Zitouni 2000] I. Zitouni (2000). *Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires : application à MAUD*. Thèse, Université Henri Poincaré, Nancy.