Recherche d'information parlée

Sommaire

Ce chapitre traite la recherche d'information à travers les thèmes complémentaires de la recherche documentaire (section 2.1) et du résumé automatique (section 2.2). Les approches présentées sont étudiées dans le cadre d'un média textuel, puis selon un média audio, en mettant l'accent sur les adaptations induites par la parole.

2.1 Recherche documentaire

La notion de recherche d'information (*information retrieval*), introduite pour la première fois par Mooers (1950), a tout d'abord été l'apanage de documentalistes ayant besoin d'un classement efficace de leurs ouvrages. Leur but était d'étendre la notion d'index, présente dans les livres, à une bibliothèque entière. Le concept d'index a été inventé, dès 1230, lorsque Hugo de St. Cher employa 500 moines pour créer une *concordance* de la Bible (Wheatley, 1879). La recherche d'information est donc née de l'exploitation du contenu d'un document pour le retrouver ; cette tâche est connue sous

le nom de recherche documentaire. Par opposition, la classification décimale de Dewey (1876) permet de retrouver des documents grâce à des méta-informations externes à l'ouvrage, selon une annotation réalisée par le documentaliste. La spécificité de la recherche documentaire est de ne réaliser qu'une partie du travail en ne présentant comme résultat non pas l'information en elle-même, mais son interprétation au sein d'un document. Le besoin de l'utilisateur est exprimé sous la forme « J'aimerais tous les documents qui parlent de ... ; je les lirai tous afin de me forger une idée exhaustive de ce sujet ». L'utilisateur est assimilé à un documentaliste recherchant non pas une information précise, mais demandant à acquérir des connaissances sur un thème donné. Cette vision du problème de la recherche d'information a l'avantage de ne nécessiter qu'une formalisation précaire des thèmes abordés dans un ouvrage : un index à base de mots devrait suffire. Pour cela, les premiers modèles de recherche d'information ont suivi un schéma simple : transformer le besoin de l'utilisateur en une série de mots-clés, puis générer la liste des documents dont l'index contient ces mots-clés. Cette approche est fonctionnelle lorsque le nombre de documents retrouvés est limité et lorsque les mots-clés choisis ne mènent pas à un trop grand nombre de documents hors-sujet (pour des raisons de polysémie). En effet, ces deux cas augmentent le temps que l'utilisateur passe à explorer les documents sans forcément obtenir une réponse à son besoin.

Pour remédier à perte de temps, il faut abandonner la problématique documentaliste et faire une étude plus approfondie du contenu des documents. Tout d'abord, les documents hors-sujet peuvent être écartés en générant non pas un ensemble de documents, mais une liste ordonnée par pertinence estimée en fonction du besoin de l'utilisateur (en comptant par exemple le nombre d'occurrences dans un document des mots-clés utilisés pour retrouver les documents). Les documents au début de cette liste sont censés contenir des informations plus intéressantes pour l'utilisateur et devraient être explorés en premier lieu. Dans un second temps, le contenu de chaque document peut être résumé en fonction du besoin utilisateur pour lui permettre de juger rapidement du potentiel informatif de ce document (en présentant par exemple le contexte d'utilisation des mots-clés déduits du besoin utilisateur). La dernière idée est de s'affranchir du document et de répondre directement au besoin de l'utilisateur, en donnant une réponse exacte à la question qu'il se pose (problématique Questions-Réponses décrite et évaluée par Voorhees, 2003). Cette notion se rapproche beaucoup plus du sens premier de la recherche d'information, mais ce domaine très intéressant demande une analyse approfondie des questions et de leurs réponses potentielles. Toutefois, elle n'est traitée relativement efficacement que pour des questions fermées ou factuelles dont la réponse est une ou plusieurs entités ou quantités (Qui ont été les présidents des États-Unis ? Combien d'habitants la France compte-elle ? ...). Les questions non factuelles du type *pourquoi* et *comment* demandent des développements construits, approchés actuellement par le résumé de documents multiples guidé par un besoin utilisateur (voir section 2.2 sur ce sujet). Il faut tout de même noter que toutes les approches pour la recherche d'information sont construites autour d'une base de connaissances (corpus, bibliothèque, base de données) constituant la Vérité et contraignant toute réponse. Bien que le raisonnement par inférence (Raina et al., 2005) puisse donner des réponses à des questions non traitées dans le socle de connaissances exploité, certaines questions

métaphysiques n'auront certainement jamais de réponse fondée de la part d'un système informatique (Il paraîtrait qu'un ordinateur ne peut répondre que 42 à la question « Quel est le sens de la vie ? », Adams, 1979).

Les problématiques de la recherche d'information sont avant tout de représenter les informations et de déduire celles qui correspondent au besoin de l'utilisateur. Mais il ne faut pas oublier que l'information est conservée sur un support dont elle doit être extraite. De plus, l'expression du besoin de l'utilisateur se fait généralement en langue naturelle. Cependant, ce besoin peut prendre d'autres formes et se retrouver étroitement lié aux résultats de la recherche d'information. Dans ce cas, l'évolution du besoin doit être analysée au travers de son reflet dans les interactions entre l'utilisateur et le système. Ce type d'analyse est primordial pour mieux estimer le besoin de l'utilisateur. Un autre problème lié à la recherche d'information réside dans la quantité de données traitées, car cette dernière impose des contraintes sur l'ensemble des problématiques précédentes (Callan, 2000).

2.1.1 Définition de la tâche

La tâche la plus répandue en recherche d'information est la recherche documentaire (*Document Retrieval*). Dans ce cadre, les informations sont matérialisées sous forme de *documents* dans une ou plusieurs modalités. Un ensemble de *documents* est appelé *corpus* et la tâche consiste à extraire d'un corpus l'ensemble des documents correspondant au besoin de l'utilisateur, exprimé sous forme d'une *requête*. La tâche est définie de façon à rendre possible une répétition des résultats car un système doit se comporter de façon déterministe dans des conditions fixées à l'avance. Historiquement, les documents et les requêtes ont été d'abord textuels, puis différents médias ont été pris en compte (son, image, vidéo). Afin de trouver les documents répondant au besoin de l'utilisateur, la plupart des approches font une analyse du contenu des documents et de la requête. L'étude de ce contenu met en jeu l'extraction d'*unités informatives* (ou *descripteurs*), le support observable de l'information. Les *unités informatives* les plus évidentes sont les mots pour un contenu textuel, les histogrammes de couleurs pour une image et les fréquences pour un signal sonore. Cette notion d'*unité informative* est dérivée du processus de généralisation, ou conceptualisation, propre au système cognitif humain. Elle implique une hypothèse d'existence de motifs représentant une même idée, une même classe d'objets, un même concept sémantique. Smoliar et al. (1996) nomment *expressives* les approches fondées sur des *unités informatives* proches des données observées et *sémantiques* les approches réalisant une analyse poussée du contenu. Nous nous intéressons dans cette partie uniquement aux unités informatives issues d'une analyse du contenu linguistique, dans l'optique d'analyser la parole extraite de documents audio.

La recherche d'information textuelle repose sur la capacité à représenter le fond (niveau sémantique) de façon indépendante de la forme (niveau syntaxique), puis d'effectuer des opérations de comparaison dans l'espace de représentation ainsi formé. Cette opération est nécessaire car la langue offre de nombreuses façons d'exprimer une idée et montre une forte variabilité de forme. Il n'existe pas de bijection entre les mots et les sens associés, un mot pouvant avoir plusieurs sens (polysémie) et plusieurs mots

pouvant avoir le même sens (synonymie). En fait, de nombreuses relations lient les concepts dénotés par les mots, comme la relation de généralisation (hyperonymie), ou de spécialisation (hyponymie). De plus, des mots peuvent agir comme représentants d'autres mots, afin d'alléger le discours. Les pronoms sont un bon exemple d'utilisation d'une forme plus courte pour faire référence à un objet que seul le contexte peut définir. Cette utilisation de plusieurs formes pour représenter un même objet ou une même idée s'appelle une anaphore ou cataphore grammaticale (à ne pas confondre avec l'anaphore rhétorique) et le phénomène est connu sous le nom de coréférence. À un plus haut niveau, les nombreuses figures de style, comme la métaphore ou l'euphémisme, altèrent le sens en offrant plusieurs niveaux d'interprétation dépendant du contexte et de la culture. Les nombreux modèles de recherche d'information essaient tous de traiter ces phénomènes de façon plus ou moins implicite, en prenant pour hypothèse qu'un champ lexical donné caractérise suffisamment bien le contenu sémantique associé. Toutefois, de plus en plus d'approches associent à ces modèles des pré-traitements linguistiques pour détecter ces phénomènes de variabilité de la forme et retrouver le fond sous-jacent.

Cette section commence par une description de la tâche de recherche documentaire et de son évaluation. Puis, les pré-traitements linguistiques les plus courants sont abordés. Ensuite, les principaux modèles pour estimer la pertinence d'un document à une requête sont présentés. L'expansion de requête vient compléter ces modèles. Enfin, l'impact de l'ensemble des méthodes précédentes sur un média parlé et un aperçu des interactions avec l'utilisateur dans ce cadre sont étudiés.

2.1.2 Évaluation

Plusieurs campagnes d'évaluation sont organisées chaque année afin de suivre les avancées dans le domaine de la recherche d'information. Ces campagnes fournissent un protocole et des données d'évaluation aux participants et réalisent un jugement de leurs performances objectif et indépendant. Les plus importantes sont *Text REtrieval Conference* (TREC[1], Voorhees et Harman, 1999), *Cross-Language Evaluation Forum* (CLEF[2], Braschler et Peters, 2004), *NII Test Collection for IR Systems* (NTCIR[3], Kando, 2005). Ces campagnes évaluent la qualité des systèmes de recherche d'information sous diverses conditions (tâche, média, langue, quantité...), selon une souche commune. Pour un besoin utilisateur donné (requête), un système doit générer une liste de réponses (documents) ordonnées par pertinence estimée. Les documents ayant le meilleur score sont considérés comme les plus susceptibles de répondre au besoin utilisateur. Dans le cadre de la recherche documentaire, les références sont constituées d'une annotation binaire[4] (pertinent / non-pertinent) de chaque document du corpus pour chaque requête évaluée. Les mesures d'évaluation utilisent la répartition entre documents pertinents et non-pertinents à un rang donné de la liste de résultats (figure 2.1).

[1] http://trec.nist.gov/, visité en octobre 2006.
[2] http://clef.iei.pi.cnr.it/, visité en octobre 2006.
[3] http://research.nii.ac.jp/ntcir/index-en.html, visité en octobre 2006.
[4] Certaines évaluations ajoutent une troisième classe pour les documents partiellement pertinents.

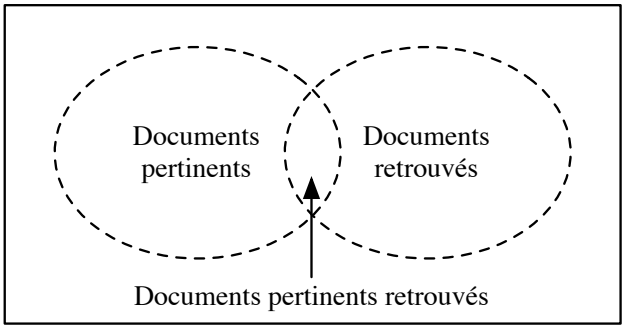


FIG. 2.1: *Illustration des différents ensembles de documents utilisés pour l'évaluation de la recherche d'information. Les mesures d'évaluations comme le rappel et la précision sont basées sur ces ensembles.*

$$R = \frac{\text{nb}(\{retrouvés\} \cap \{pertinents\})}{\text{nb}(\{pertinents\})} \tag{2.1}$$

$$P = \frac{\text{nb}(\{retrouvés\} \cap \{pertinents\})}{\text{nb}(\{retrouvés\})} \tag{2.2}$$

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{2.3}$$

La précision (P, équation 2.2) est le nombre de documents pertinents retrouvés par rapport au nombre de documents retrouvés, alors que le rappel (R, équation 2.1) est le nombre de documents pertinents retrouvés par rapport au nombre de documents pertinents du corpus. La précision est tracée en fonction du rappel afin de comparer ces deux mesures au sein d'une courbe (courbe de précision-rappel dont un exemple est donné par la figure 2.2). Un score général d'un système sera donné par la précision moyenne sur onze points de rappel $(0.0, 0.1..., 1.0)$, sur l'ensemble des requêtes (*Mean Average Precision*, MAP). Une autre façon de combiner précision et rappel est d'utiliser la F_1-mesure, moyenne harmonique de P et R (équation 2.3, avec $\beta = 1$). Ces mesures d'évaluation sont focalisées sur une tâche répondant à des documentalistes car elles prennent en compte un parcours exhaustif des documents. Afin de rendre compte d'une utilisation plus actuelle des moteurs de recherche, il faut employer une mesure intégrant le fait que les utilisateurs ne parcourent que les deux ou trois premières pages des résultats avant d'estimer que l'information recherchée ne sera pas retrouvée. La précision à n résultats (avec $n < 20$) est une bonne estimation de la qualité de la réponse au besoin de l'utilisateur, en ignorant toutefois la difficulté de la requête (Carmel et al., 2005).

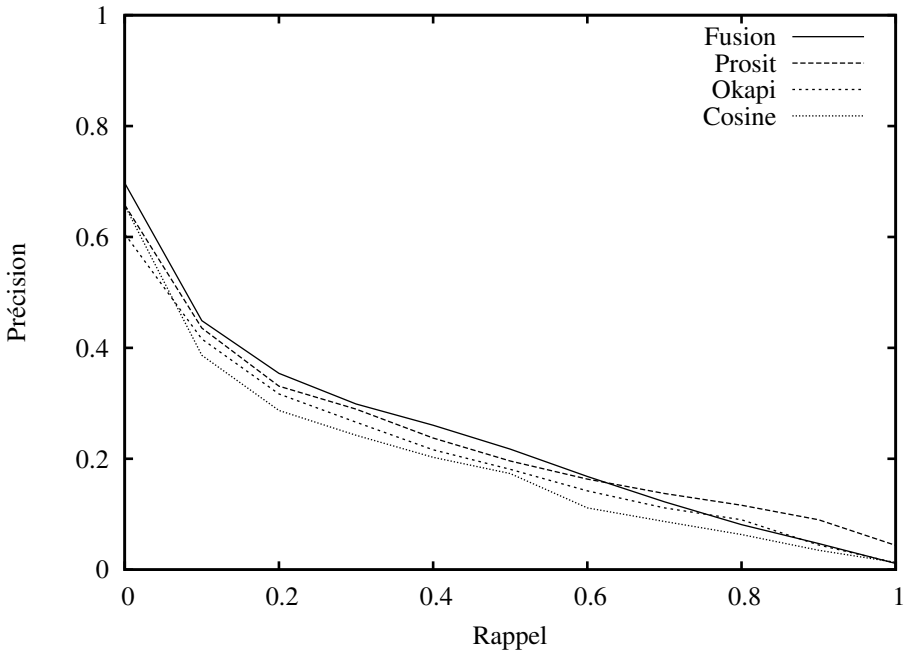


FIG. 2.2: *Exemple de comparaison de systèmes en recherche documentaire à l'aide de courbes de précision-rappel sur la campagne TREC-8. Les systèmes présentés sont le modèle vectoriel classique (Cosine), un modèle fonction de la quantité d'information (Prosit), un modèle probabiliste (Okapi), et la moyenne pour chaque document des scores normalisés des trois modèles précédents (Fusion). Ces différents systèmes suivent les formulations proposées par Savoy et Berger (2005).*

2.1.3 Pré-traitements linguistiques

Des pré-traitements linguistiques sont appliqués avant la transposition des documents dans un espace sémantique dans le but de minimiser l'impact des ambiguïtés. Lorsque le média traité est le texte, une première étape consiste à normaliser la forme de surface (correction orthographique, réaccentuation, normalisation de la casse, expansion des abréviations, normalisation des valeurs numériques, suppression du contenu non informatif...). Ce premier traitement correspond à un nettoyage des données, généralement implémenté à l'aide de règles de réécriture et de lexiques. Après ce traitement, les mots outils (déterminants, conjonctions, certains adverbes...) sont identifiés et supprimés car ils ne sont pas discriminants pour représenter le thème d'un document.

Les différents mots-formes d'un texte peuvent être regroupés sur la base de leur lemme, afin d'éliminer une partie de la variabilité morphologique et de créer des représentants plus « robustes » des concepts. Une première technique, appelée lemmatisation, remplace les mots par leur forme canonique (indépendante du genre, du nombre, de la personne et du mode). En général, elle est appliquée après étiquetage morpho-

syntaxique à l'aide d'un dictionnaire de triplets (*mot, étiquette, lemme*). Une autre technique cherche à réduire les mots à leur racine de façon algorithmique : la *racinisation* (Porter, 1980) est une approche heuristique dépendante de la langue et contenant un risque lié à de nombreux cas particuliers[5] créant des erreurs de confusion.

Toutefois, l'approche précédente ne remplace pas une désambiguïsation des mots. Par désambiguïsation, il est suggéré que le sens d'un mot peut être découvert parmi ses différents sens possibles, en utilisant le contexte autour de ce mot. Les différentes techniques pour la désambiguïsation[6] sémantique automatique n'ont pas de très bonnes performances dans le cas général (Agirre et Edmonds, 2006). Connaître le sens dans lequel les mots ont été employés permet de les projeter dans une ressource sémantique décrivant leurs relations (synonymie, antonymie, hyperonymie, hyponymie...), comme Wordnet (Miller, 1995).

Afin d'aller plus loin, des entités informatives liées au domaine peuvent être annotées (détection des entités nommées, sujet traité dans la section 4.2), les expressions anaphoriques pronominales peuvent être résolues (Lappin et Leass, 1994) et les figures de style peuvent être reconnues (Markert et Nissim, 2002). Ces domaines florissants tentent de résoudre des problèmes difficiles mais il est évident qu'ils auront de plus en plus d'applications en recherche d'information.

Les approches pour annoter ces phénomènes reposent souvent sur une analyse statistique des propriétés des objets en jeu, et des algorithmes d'apprentissage capables de reconnaître des classes d'objets en fonction de leurs caractéristiques. Les gains que peuvent apporter ces approches sur une tâche donnée sont très dépendants du domaine traité et des corpus utilisés. Ces gains ne sont pas toujours cumulatifs en raison de l'introduction d'erreurs ou d'ambiguïtés dans les différents pré-traitements. Dans l'optique de traiter des documents audio, il serait souhaitable d'appliquer ces techniques afin d'améliorer la perception du message parlé. Les traitements de plus haut niveau, ayant déjà de faibles performances sur le texte, risquent de se dégrader à cause de la variabilité de la parole.

2.1.4 Modèles

La plupart des modèles de recherche d'information sont soit définis pour une tâche de recherche documentaire (comparaison d'une requête à un document), soit pour une tâche de catégorisation (comparaison d'un document avec un autre document, ou avec un pseudo-document). Bien que ces deux types de tâches impliquent des hypothèses différentes lors de la modélisation, elles sont souvent considérées comme équivalentes dans la pratique, en considérant un document comme étant la requête, généralement au détriment de légers changements dans le comportement des modèles. De plus, les modèles se réfèrent aux *unités informatives* observées sous différents noms (mots, termes,

[5] Par exemple, le *stemmer* Snowball (`http://snowball.tartarus.org`, visité en décembre 2006) pour le français fait correspondre « guérir » et « guerre » à travers le radical « guer ».

[6] Plus d'informations peuvent être obtenues sur `http://www.senseval.org`, campagne d'évaluation des systèmes de désambiguïsation sémantique, visité en octobre 2006.

concepts...) selon les pré-traitements appliqués. Nous présentons dans cette section les modèles les plus répandus, dont une taxonomie est donnée par la table 2.1. Des ouvrages comme (Baeza-Yates et Ribeiro-Neto, 1999) ou (Ihadjadene, 2004) détaillent les formulations des modèles présentés dans cette section.

Type / **Cadre**	**Sans interdépendance**	**Interdépendance**	
		Intrinsèque	*Extrinsèque*
Ensembliste	Booléen Booléen étendu		Ensembles flous
Algébrique	VSM	GVSM Réseaux de neurones LSA *Random Indexing*	TVSM Infomap
Probabiliste	BIR Réseaux bayesiens	Modèles de langage Modèles de pertinence pLSA, LDA	RbI

TAB. 2.1: *Taxonomie des principaux modèles en recherche d'information inspirée par Kuropka (2004), en fonction de leurs propriétés et de leur origine mathématique. Ces modèles reposent sur des hypothèses fortes liées au degré de dépendance entre les unités de surfaces observées (mots, termes ou concepts), et la façon de modéliser cette dépendance (intrinsèque ou extrinsèque). Les acronymes sont détaillés tout au long de la section.*

Modèles ensemblistes

Le modèle booléen s'appuie sur la logique de Boole en permettant à l'utilisateur d'exprimer son besoin par une formule logique sur les mots. Les documents satisfaisant cette formule sont considérés comme pertinents et renvoyés à l'utilisateur. Lorsque cette satisfaction est binaire, il n'y a pas de notion de degré de pertinence, donc le modèle n'est utile que dans les cas où le nombre de résultats retrouvés correspond aux attentes de l'utilisateur. Afin d'y remédier, le modèle booléen a été étendu de nombreuses manières. Par exemple, Salton et al. (1983) ajoutent un degré de satisfaction de la requête logique fonction des propriétés statistiques des mots et une pondération des opérateurs logiques « ET » et « OU ». Ces approches s'apparentent autant au modèle booléen qu'au modèle vectoriel. Parallèlement aux modèles fondés sur la théorie des ensembles « stricts », il existe d'autres modèles basés sur la théorie des ensembles flous qui prennent en compte la corrélation entre les mots pour définir l'appartenance de chaque document aux ensembles des mots qui les composent. Les performances de cette classe de modèles ont rarement été comparées à celles des modèles les plus répandus.

Modèles algébriques

Les modèles algébriques utilisent une projection des documents et des requêtes dans un espace vectoriel dans lequel ces vecteurs peuvent être comparés. La pertinence

estimée d'un document est directement proportionnelle à une mesure de sa similarité à la requête. Le modèle vectoriel (*Vector Space Model*, VSM, Salton et al., 1975) est le modèle le plus populaire en recherche d'information car il permet d'obtenir des performances intéressantes en nécessitant peu de ressources. L'idée de cette modélisation est d'exprimer chaque documents ($\vec{d}$) et la requête ($\vec{q}$) comme des vecteurs dans l'espace formé par le vocabulaire (équations 2.4 et 2.5). Chaque dimension de cet espace représente un mot du vocabulaire (u_i) ; la composante du vecteur document (ou requête) pour ce mot, $w_{i,d}$ est donnée dans l'équation 2.6, en fonction de la fréquence du mot dans le document, $f_i(d)$, par rapport à sa fréquence dans le corpus (N est le nombre de documents, et n_i le nombre de documents où apparaît le mot u_i). Cette normalisation par rapport à la fréquence dans le corpus est appelée *Inverse Document Frequency* (IDF) et permet de spécifier que les événements peu fréquents sont plus susceptibles d'intéresser l'utilisateur. Cette notion est présente dans la recherche d'information depuis ses débuts (Bar-Hillel, 1958), mais a fait l'objet de nombreuses formulations. A partir de la représentation vectorielle d'un document, sa pertinence à la requête est estimée en calculant sa similarité avec le vecteur requête. La similarité la plus répandue est la similarité *cosine*, cosinus de l'angle entre les deux vecteurs $\vec{q}$ et $\vec{d}$, comme le détaille l'équation 2.7. Savoy et Berger (2005) comparent les formulations de pondération pour tf et idf et les différentes similarités sur la campagne CLEF 2005. Certaines par exemple prennent en compte le biais de la longueur des documents et de la fréquence de mots intra-document pour permettre des améliorations significatives au sens de l'évaluation.

$$\vec{d} = (w_{0,d}..., w_{|W|,d})^T \tag{2.4}$$

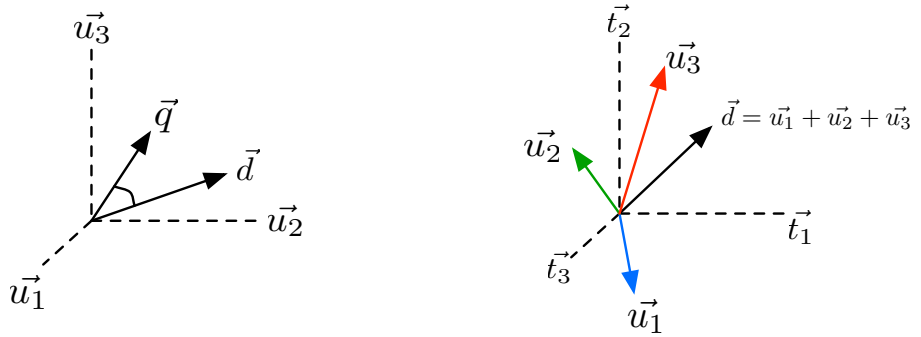
$$\vec{q} = (w_{0,q}..., w_{|W|,q})^T \tag{2.5}$$

$$w_{i,d} = tf_{i,d} \times idf_i = log(1 + f_i(d)) \times \log \frac{N}{n_i} \tag{2.6}$$

$$cosine(\vec{q}, \vec{d}) = \frac{\vec{d} \cdot \vec{q}}{|\vec{d}| \times |\vec{q}|} = \frac{\sum_{i=0}^{|W|} w_{i,d} w_{i,q}}{\sqrt{\sum_{i=0}^{|W|} w_{i,d}^2} \sqrt{\sum_{i=0}^{|W|} w_{i,q}^2}} \tag{2.7}$$

Malgré la faible complexité du calcul de la similarité entre deux documents qui rend celui-ci adapté aux grands corpus, il est souvent reproché à ce modèle de ne pas prendre en compte l'ordre des mots (modèles à sac-de-mots), ni même la relation entre les mots (« maison » et « blanche » sont des mots très répandus, donc faiblement pondérés alors que « maison blanche » devrait avoir un impact beaucoup plus fort sur la similarité).

Wong et al. (1985) ont proposé le modèle vectoriel généralisé (*Generalized Vector Space Model*, GVSM) pour prendre en compte les corrélations inter-mots. Le modèle vectoriel impose une base orthonormale de l'espace des documents. Cette base implique que chaque vecteur représentant un mot est orthogonal à tous les autres vecteurs représentant des mots. Dans le modèle vectoriel généralisé, le vecteur représentant un mot est défini selon sa corrélation avec les autres mots du lexique. Ainsi, comme l'illustre la figure 2.3, un vecteur document (somme des vecteurs représentant les mots qu'il contient) prendra en compte les affinités des mots à apparaître ensemble. La complexité de ce modèle est beaucoup plus grande que celle du modèle classique, pour un gain de



modèle vectoriel classique (VSM) modèle vectoriel généralisé (GVSM)

FIG. 2.3: *Illustration de la différence entre le modèle vectoriel classique et le modèle vectoriel généralisé et autres modèles impliquant des vecteurs mots non orthogonaux. $\vec{q}$, $\vec{d}$, $\vec{u_i}$ et $\vec{t_j}$ représentent respectivement la requête, un document, un mot (unité informative) et une dimension sous-jacente aux mots (thème).*

performance pas toujours convaincant (augmentation du rappel, diminution de la précision) .

Latent Semantic Analysis (LSA), également référencé sous le nom de *Latent Semantic Indexing* (LSI), peut être vu comme une extension de GVSM offrant une réduction de la taille de l'espace de comparaison et donc de la complexité (Deerwester et al., 1990). La méthode repose sur une réduction de la matrice documents-mots X à ses dimensions principales en utilisant une décomposition en valeurs singulières (SVD). La matrice est décomposée en une multiplication de la matrice de vecteurs singuliers gauche U, de la matrice diagonale de valeurs singulières Σ et de la matrice de vecteurs singuliers droite V^T (équation 2.8). Lorsque les valeurs singulières sont ordonnées de la plus grande à la plus petite, réduire le rang k de la matrice Σ correspond à approximer la matrice X en minimisant l'erreur au sens de la norme L^2 entre les mots (équation 2.9, dans laquelle $\hat{X}$ est l'approximation de X et Σ_k la matrice de valeurs propres réduite au rang k). De plus, les dimensions de l'espace réduit font apparaître des « thématiques » selon lesquelles sont exprimés les documents. Une projection de la requête dans cet espace permet de calculer une similarité prenant en compte les caractéristiques thématiques de la requête. Ce modèle est plus largement décrit dans la section 5.3 où il est appliqué au résumé automatique orienté par une requête.

$$X = U\Sigma V^T \tag{2.8}$$

$$\hat{X} = U\Sigma_k V^T \tag{2.9}$$

LSA offre une réduction de dimension de bonne qualité, mais nécessite d'évaluer la matrice d'occurrences entre mots et documents et de décomposer cette matrice. Afin d'éviter cette réduction coûteuse en ressources, Kanerva et al. (2000) proposent une technique nommée *Random Indexing*. Cette approche consiste à associer aux mots des vecteurs aléatoires quasi-orthogonaux (contenant un grand nombre de 0 et un petit nombre de -1 et $+1$), de dimension fixe et de construire un équivalent de la matrice

de coocurrences grâce à des accumulateurs. La réduction est d'aussi bonne qualité que pour LSA, avec l'avantage de supporter le passage à l'échelle et de pouvoir être mise à jour de façon incrémentale. De nombreux autres modèles algébriques sont décrits dans la littérature tels que la recherche documentaire à base de réseaux de neurones (Wilkinson et Hingston, 1991), utilisant les mots de la requête en entrée et générant les documents en sortie ; l'emploi de LSA pour créer les vecteurs de mots du modèle vectoriel généralisé (Infomap, voir section 5.3) ; et les modèles vectoriels thématiques (TVSM, Becker et Kuropka, 2003) associés à une classification en thèmes des mots et/ou des documents.

Modèles probabilistes

Le cadre probabiliste est très attirant pour la recherche documentaire et a été appliqué de nombreuses façons. Trois événements entrent en jeu dans ce cadre : la requête Q, un document D et la pertinence binaire R. Les modèles proposés diffèrent par leur estimation de la probabilité qu'un document soit pertinent pour une requête donnée $P(R = 1|D, Q)$. La figure 2.4 illustre les principales approches, à savoir : le modèle classique, les modèles de langage et les modèles de pertinence. Le modèle classique de Robertson et Spärck-Jones (1988) est fondé sur le ratio de vraisemblance entre $P(R = 1|D, Q)$ et $P(R = 0|D, Q)$, les modèles de langage de Ponte et Croft (1998) estiment la probabilité que la requête soit issue de la même distribution qu'un document $P(Q|D)$ et les modèles de pertinence (Lavrenko, 2002) utilisent la divergence entre les distributions $P(Q|R)$ et $P(D|R)$. Les probabilités incluant la pertinence sont difficiles à estimer dans le cas où cette variable n'est pas observée (schéma de recherche d'information classique). Par contre, lorsqu'un *a priori* sur la pertinence est fourni (par exemple grâce à des interactions utilisateur), ces modèles ont un avantage certain.

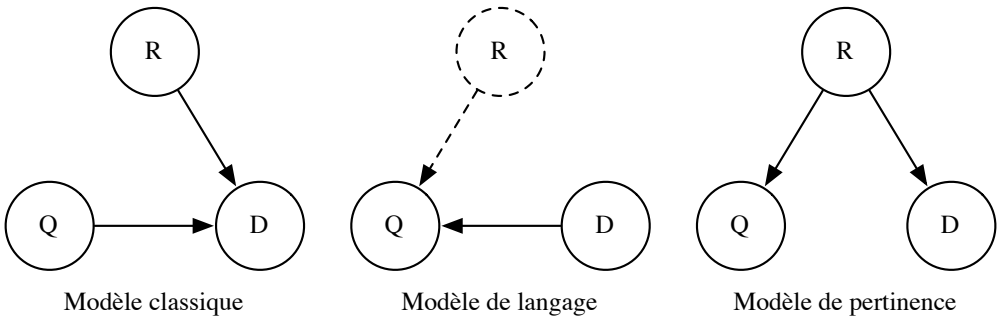


FIG. 2.4: *Différentes modélisations probabilistes pour la recherche documentaire selon le formalisme des modèles graphiques. Ces modèles estiment la pertinence d'un document pour la requête en fonction des événements Q (la requête), D (le document) et R (la pertinence). Dans cette figure, une absence de flèche correspond à une hypothèse d'indépendance entre deux événements. Les éléments en pointillés sont implicites au modèle et n'apparaissent pas dans sa formulation.*

Le modèle classique, nommé *Binary Independence Retrieval* (BIR), fait l'hypothèse que les mots sont indépendants deux à deux. Ses dernières extensions sont détaillées dans

(Spärck-Jones et al., 2000). La probabilité qu'un document soit pertinent est exprimée comme le produit des probabilités de pertinence de chaque mot qui le compose. Cette probabilité est modélisée par une distribution 2-poisson prenant en compte l'« élitisme » d'un mot pour un document (sur-pondération des mots très représentatifs du thème) en fonction de son nombre d'occurrences (Robertson et Walker, 1994).

Les modèles de langage (Ponte et Croft, 1998) tentent d'introduire une meilleure estimation de la distribution des mots dans les documents la modélisant par une loi multinomiale. Toutefois, Church et Gale (1995) prouvent que ce type de distribution ne reflète pas complètement la réalité. En général, le cadre du maximum de vraisemblance est utilisé pour estimer la probabilité pour un modèle de document de générer la requête. Les idées du modèle algébrique LSI/LSA ont été reprises par pLSI/pLSA, une approche probabiliste de l'analyse en sémantique latente (Hofmann, 2000). Cette approche définit un modèle génératif de la matrice de comptes selon lequel les mots proviennent de k thèmes mis en relation par des variables latentes de mélange. pLSI a le désavantage de nécessiter l'estimation de beaucoup de paramètres et d'aboutir à du sur-apprentissage. Pour lever ces problèmes, Blei et al. (2003) proposent *Latent Dirichlet Allocation* (LDA), un modèle où les mots des documents sont toujours générés par une distribution multinomiale. Néanmoins, les thèmes latents sont considérés comme des variables régissant le comportement des paramètres de la multinomiale à travers une distribution de Dirichlet. Ce modèle implique une estimation coûteuse de ses paramètres (généralement approximée) sans toutefois apporter les gains en performance escomptés sur une tâche de recherche documentaire (Wei et Croft, 2006).

Il existe d'autres modèles probabilistes fondés sur une topologie de réseaux bayesiens, comme par exemple (Callan et al., 1992), à la base du système Inquery. Ce modèle cherche à estimer l'inférence $d \rightarrow u \rightarrow q$. Cette inférence est aussi estimée dans le cadre plus général des réseaux de croyances (Ribeiro-Neto et Muntz, 1996). Une autre méthode, *Retrieval by Logical Imaging* (RbI, Crestani et van Rijsbergen, 1995) apporte une vision intéressante en utilisant une logique modale pour étendre à un cadre probabiliste les idées du modèle booléen (*Retrieval by Logical Imaging*, RbI).

La prochaine section est dédiée à l'expansion de requête, un processus indispensable pour les modèles probabilistes.

2.1.5 Expansion de requête

Une requête est censée représenter le besoin de l'utilisateur. Or, la langue naturelle offre de nombreuses possibilités de varier la forme de surface tout en conservant un contenu sémantique proche. L'expansion de requête consiste à améliorer la représentation du besoin utilisateur grâce aux interactions issues d'une première recherche d'information. Par exemple, l'utilisateur peut explicitement désigner parmi les résultats des documents pertinents et des documents non-pertinents, ce qui peut être utilisé pour favoriser (resp. défavoriser) les mots apparaissant dans ces documents. Pour prendre un autre exemple, un moteur de recherche offrant un aperçu des documents retrouvés pourra tirer avantage de l'ordre dans lequel l'utilisateur visualise les documents. Les

modèles de recherche documentaire doivent intégrer ces informations pour améliorer les prochaines recherches, au travers d'un processus d'expansion de requête.

Dans le cadre des modèles vectoriels, les mots des documents désignés comme pertinents sont ajoutés avec un facteur positif pour la requête, alors que les mots des documents désignés non-pertinents sont ajoutés avec un facteur négatif (Buckley et al., 1994). Ce processus permet d'améliorer efficacement les performances de la recherche. Les modèles probabilistes qui intègrent la notion de pertinence ont alors une observation directe des probabilités $P(R|D)$ et $P(\overline{R}|D)$, qui évite d'utiliser des approximations de mauvaise qualité pour ces probabilités.

Cette utilisation est limitée aux scénarios incluant des utilisateurs. Il est heureusement possible de réaliser une expansion en aveugle en estimant que les documents en haut du classement ont une plus grande probabilité d'être pertinents que ceux en bas du classement. Dans la pratique, les n premiers documents sont considérés comme pertinents. Ce processus peut être itéré, mais il n'a jamais été prouvé qu'il convergeait vers les documents pertinents. Dans le meilleur des cas, un petit nombre d'itérations permet de fortement améliorer le rappel, au détriment de la précision (ce qui correspond à relâcher les contraintes désignant les documents recherchés).

D'autres types d'expansion ont été proposés, comme l'utilisation d'ontologies pour augmenter les mots de la requête avec leurs synonymes (Gonzalo et al., 1998), ou encore étendre le champ lexical d'un document à l'aide de ses voisins dans un corpus « propre » de grande taille (Singhal et Pereira, 1999) .

L'expansion de requête et de document fonctionne bien pour augmenter le rappel, mais la précision peut être améliorée en effectuant un partitionnement non supervisé des meilleurs documents en thèmes, et en supprimant les partitions de petite taille correspondant à des résultats aberrants (de Loupy et al., 1998). Une autre possibilité pour augmenter la précision consiste en la fusion de plusieurs listes de résultats générées selon différentes combinaisons de pré-traitements, de modèles et de paramètres (Bellot, 2000). La moyenne, le minimum ou le maximum des classements de chaque document sont des approches simples dans le cas où aucun *a priori* n'est connu sur la qualité des classements (Bellot et El-Bèze, 2000). Dans le cas contraire, il est possible d'apprendre les paramètres du reclassement en fonction de la source, de l'espace des scores ou à l'aide de propriétés des documents et des requêtes (Croft, 2000).

2.1.6 Extension à la parole

Les premières approches de la recherche d'information dans un contenu parlé ont d'abord utilisé des techniques similaires à celle développées pour les documents textuels, appliquées à la transcription automatique du flux de parole. La recherche documentaire audio (*Spoken Document Retrieval*, SDR) est la première formalisation de la tâche au travers de la campagne TREC 7. Cette tâche est associée à la recherche d'information dans des documents papier numérisés par *Optical Caracter Recognition* (OCR) car, dans les deux cas, les erreurs introduites peuvent être assimilées à un bruitage du contenu linguistique originel. La tâche SDR de TREC (Garofolo et al., 1999) consiste à

indexer 500 heures d'émissions radio en anglais, en utilisant les transcriptions automatiques (à différents taux d'erreur de mots, *Word Error Rate*, WER) des documents issus de la campagne Hub 4 (Przybocki et al., 1998) organisée par NIST. L'information recherchée dans les documents audio est exprimée sous la forme d'une requête textuelle semblable à celles exploitées dans TREC *ad-hoc*. Pour plus d'informations, le lecteur pourra se référer à l'atelier spécial sur l'adaptation des techniques de recherche documentaire aux applications impliquant de la parole (Coden et al., 2002), organisé lors de la conférence SIGIR 2001. Il faut remarquer que la plupart des systèmes de recherche documentaire fonctionnent soit sur du texte soit sur l'audio, mais ne mélangent pas les deux modalités. Sanderson et Shou (2002); Favre (2003) soulignent qu'en général ce mélange défavorise l'audio et qu'aucune technique permettant de réduire cet écart n'a été proposée à ce jour.

Les évaluations TREC montrent que le taux d'erreur de mots est linéairement corrélé aux performances en recherche documentaire et qu'un taux d'erreur inférieur à 40% permet d'obtenir des résultats acceptables par l'utilisateur (Garofolo et al., 1999). Cette bonne réussite s'explique d'abord par la longueur des requêtes TREC et la quantité d'informations qu'elles contiennent (environ 10 mots porteurs de sens, à comparer à des requêtes WEB de moins de 2 mots en moyenne). L'impact du taux d'erreur peut être limité à 10% des performances sur la transcription manuelle en utilisant des techniques d'expansion de requête et de document comme celles présentées dans la section 2.1.5. Johnson et al. (2000) notent que le gain des différentes techniques n'est pas cumulatif et que l'utilisation de corpus externes propres (et thématiquement proches des données traitées) pour l'expansion est bénéfique. Toutefois, Hansen et al. (2004) font face à des conditions moins favorables sur les données de la *National Gallery of the Spoken Word* (NGSW) avec des taux d'erreur de mots de 40% et observent qu'un bon choix des paramètres utilisés lors de l'expansion permet d'obtenir un cumul des gains (20% relatif au total).

Des données audio sont utilisées dans le cadre d'une autre tâche intéressante lors de la campagne *Topic Detection and Tracking* (TDT), pour laquelle il faut faire du suivi de thème et détecter les nouveautés dans le flux d'informations (Allan, 2002). Cette tâche a impliqué la mise au point de nouvelles méthodes de détection de coupures thématiques en utilisant aussi bien le contenu liguistique que le contenu audio. Les techniques de recherche d'information audio tentent maintenant d'aller plus loin que la parole des flux radio, en se focalisant sur la parole spontanée et sur les applications temps réel. Brown et al. (2001), par exemple, s'attachent à annoter des flux télévisuels et des conférences avec des informations susceptibles d'intéresser le spectateur. Pour ce qui est de la parole spontanée, Byrne et al. (2004) ont annoté un corpus de 10000 heures d'interview puisées dans les enregistrements de la *Shoah Visual History Foundation*. Ce corpus est à ce jour le plus grand corpus de parole spontanée réunissant de nombreux locuteurs sur le même thème ; ce corpus permettra certainement de mieux tester les approches de recherche d'information et de segmentation que les corpus téléphoniques de Switchboard (Godfrey et al., 1992).

Le taux d'erreur de mots n'est pas le seul problème lié à la transcription automatique du contenu parlé, les systèmes de transcription ont en effet un vocabulaire limité aux

mots les plus fréquents (dans le but de minimiser le taux d'erreur de mots, tout en limitant les ressources nécessaires). Les mots les moins fréquents sont considérés comme des mots hors vocabulaire (*Out of Vocabulary*, OOV) et ignorés lors du décodage du signal de parole. Ils ne pourront être retrouvés et paradoxalement, ce sont justement les événements peu fréquents et inattendus qui sont le plus susceptibles de sélectionner les documents pertinents. En effet, le moteur de recherche SpeechBot (Thong et al., 2000) a offert pendant plusieurs années l'accès à du contenu parlé transcrit automatiquement sur le web et il a été observé que plus de 12% des mots utilisés dans les requêtes étaient hors vocabulaire. Le problème est aussi lié aux modèles de langages nécessairement mal estimés pour les langues à ressources minoritaires comme les langues africaines (Abdillahi et al., 2006). Des techniques basées sur l'utilisation de sous-parties des mots comme les phonèmes ou les radicaux sont apparues pour essayer de remédier au problème des mots hors vocabulaire (Wechsler et al., 1998). Ces approches demandent une phonétisation de la requête, puis la comparaison de cette séquence de phonèmes avec les hypothèses de transcription phonétique du système de transcription. Une mesure de confiance basée sur l'adéquation entre la modélisation phonétique et le contenu acoustique est utilisée afin de ne rapporter que des séquences proches de la meilleure hypothèse (probabilité *a posteriori* du sous-graphe d'hypothèses passant par le chemin étudié). L'utilisation du treillis[7] de phonèmes apporte un gain intéressant en rappel au détriment de la précision car de nombreux passages ont une transcription phonétique similaire à la requête sans pour autant impliquer la présence des mêmes mots. Yu et Seide (2004) intègrent la recherche dans le treillis de phonèmes avec une recherche dans le treillis de mots afin de profiter de l'augmentation à la fois du rappel et de la précision. Face à un taux d'erreur de mots de l'ordre de 43% à 60% selon les conditions, ils observent un gain de 10% en performance sur la détection de mots (*word spotting*) par rapport à l'utilisation d'une des deux méthodes isolément. Les mots hors vocabulaire ont des effets de bord sur la qualité de la transcription, car ils sont remplacés par une séquence de mots acoustiquement proches, mais qui diverge du contenu réel et provoque des erreurs autour du mot inconnu. Bazzi et Glass (2000) proposent par exemple d'introduire un mot « INCONNU » dans le vocabulaire et d'utiliser un modèle phonétique complet pour représenter son acoustique. Cette approche par cas particulier n'entre pas dans les cadres mathématiques utilisés en transcription et demande un contrôle fin de son activation.

Nous retiendrons que la recherche d'information audio s'est surtout concentrée sur l'interaction transcription/recherche. Pour preuve, le standard MPEG 7 (Manjunath et al., 2002) adopte, entre autre, la représentation par treillis d'hypothèses phonétiques pour le stockage des transcriptions automatiques de flux structurés afin d'autoriser la remise en cause du lexique de reconnaissance.

2.1.7 Interaction avec l'utilisateur

Gilbert et Zhong (2003) observent que des interfaces efficaces pour la recherche d'information audio restent peu développées, car l'effort de recherche s'est concentré sur

[7] le mot « treillis » est utilisé dans le sens de graphe d'hypothèses, de l'anglais *lattice*.

des modèles correspondant à une application générique.

Une première interface a été proposée par Arons (1993) avec pour objectif de faciliter la navigation dans un signal sonore. Des fonctions d'avance rapide et retour rapide sont étudiées en réduisant d'abord les silences inter-mots, puis en supprimant les mots à faible intonation. Elles permettent de naviguer 5 fois plus rapidement dans le flux de parole, mais certainement pas d'aborder des bases de grande taille. Le principe de focalisation, grâce auquel l'homme est capable de suivre un locuteur au milieu de plusieurs conversations, a été exploité par Kobayashi et Schmandlt (1997) qui spatialisent le son pour diffuser plusieurs signaux en mouvement. Ainsi, l'axe temporel est projeté dans l'espace et l'utilisateur fait alors appel à sa mémoire spatiale (généralement plus développée que la mémoire temporelle). L'utilisateur peut à tout moment réécouter un segment intéressant tout en continuant la lecture du flux principal. Cette approche n'a connu que peu de succès compte tenu de l'infrastructure nécessaire pour une spatialisation efficace du son. Avec l'arrivée des méthodes de recherche d'information appliquées à la parole, des interfaces intégrant un moteur de recherche se sont développées (Hirschberg et al., 1999; Thong et al., 2000) afin de mieux cibler le besoin utilisateur. Malheureusement, il est difficile sur les grandes bases de données audio d'aller à l'essentiel et une requête simple peut engendrer une grande quantité de segments longs à écouter. Pourtant, l'utilisation de la transcription reste le moyen le plus efficace d'extraire une sémantique du signal audio. Cependant, Hirschberg et al. (2001) rapportent que les utilisateurs de SCANMail, un système de présentation de messages vocaux, ont tendance à trop faire confiance à la transcription automatique et à ne pas écouter le message réel. De plus, tous leurs efforts pour créer des présentations alternatives à la transcription (mise en valeur des segments importants sur une frise chronologique...) sont généralement inutilisés car beaucoup moins intuitifs. Dans le même domaine que la recherche documentaire textuelle, la problématique Questions-Réponses (Q/A) se transpose naturellement à l'audio sous la forme d'un dialogue homme-machine (Galibert et Rosset, 2005; Hori et al., 2003c; Varges et al., 2006; Stenchikova et al., 2006). L'avantage principal du dialogue est de pouvoir demander des précisions sur la question ou une reformulation en cas de transcription peu confiante. Par contre, il faut être capable de formuler la réponse la plus courte possible et d'utiliser les mêmes artifices que l'être humain pour faire patienter son auditeur. Le dialogue permet d'optimiser la granularité de la réponse en étant d'abord généraliste et en demandant à l'utilisateur s'il souhaite plus de détails. La problématique temps réel est impérative pour une bonne interaction et le système doit gagner du temps pour trouver sa réponse. Il peut le faire en tirant parti de la fonction phatique (ajout d'éléments lexicaux pour ne pas rompre le discours), ou par des formulations longues (choix de synonymes plus longs à prononcer). Les applications issues de Q/A sont pour l'instant surtout limitées à des questions dont la réponse directement accessible dans les documents sources, mais les approches devraient, comme ça a été le cas pour Q/A sur le texte, rejoindre le résumé automatique pour être étendues aux questions non factuelles.

2.2 Résumé automatique

Cette section présente les approches majeures[8] en résumé automatique de texte et leur extension à la parole. Le lecteur trouvera de nombreuses informations sur le sujet dans (Mani, 2001).

Historiquement, le résumé automatique a d'abord été appliqué au texte. En effet, la première approche a été proposée lorsque les premiers ordinateurs ont été capables de numériser des documents. Luhn (1958) extrait des mots-clés représentatifs du contenu d'un document à l'aide de statistiques et se sert de ces mots-clés pour sélectionner les phrases les plus importantes du document. L'idée du résumé par extraction est restée une des approches les plus répandues pendant les 50 années qui suivirent. Edmundson (1969) introduit l'idée d'utiliser les caractéristiques des phrases afin d'évaluer leur importance. Cette idée se traduit par une analyse des documents traités suggérant que les phrases importantes sont caractérisées par la présence d'indicateurs lexicaux (« Ce document décrit ... », « Pour conclure... »), par la présence de mots-clés, par le nombre de mots en commun avec le titre, et par la position de la phrase dans le document. Une combinaison de ces différents paramètres dans le choix des phrases extraites permet d'améliorer le nombre de mots en commun avec un résumé écrit à la main. Elle est aussi une des premières approches de l'évaluation de la qualité d'un résumé. Après quelques échecs de méthodes linguistiques observés par Paice (1990), Kupiec et al. (1995) formulent le problème de sélection de phrases candidates au résumé comme un problème de classification. Des paires (*document d'origine – résumé*) sont nécessaires pour apprendre les paramètres optimaux du classifieur. Parallèlement à ces méthodes statistiques, de nombreuses approches utilisent des connaissances du domaine traité pour extraire les informations importantes (McKeown et Radev, 1995). Par exemple, un résumé sur la vie d'une personne nécessite de détecter sa date de naissance, des informations sur sa famille, ses études, ses activités professionnelles...Une fois ces informations extraites, un système de génération les utilise pour synthétiser la séquence linguistique correspondant au résumé. La qualité des résumés produits est fortement dépendante de la qualité des patrons utilisés et reste très dépendante du domaine.

Le résumé automatique n'a pas été appliqué uniquement au texte. Dans le domaine de la vidéo, une bande annonce de film peut s'apparenter à un résumé dont le but est de donner envie au spectateur de voir le film. Dans ce domaine, Nam et Tewfik (1999) génèrent des résumés de séquences vidéo par extraction des zones de forte activité des modalités observées dans le temps, sans prendre en compte l'intérêt du spectateur. L'objectif d'un résumé ne doit pas être négligé. Il est impensable, par exemple, que le résumé au dos d'un livre dévoile l'intrigue. En fait, les résumés peuvent être classés selon différents critères :

1. **Le but**
   Le résumé *indicatif* permet de savoir s'il faut approfondir un sujet ; le résumé

[8] Le site http://summarization.com, visité en octobre 2006, est une bonne source d'informations sur le résumé automatique de texte. Maintenu par D. Radev, il propose par ailleurs une liste de 700 références bibliographiques dans le domaine.

informatif donne une information généraliste et objective ; le résumé *critique* donne le point de vue de son auteur.

2. **La forme**
   Un résumé peut prendre la forme d'une succession d'*extraits* représentatifs, ou consister en une formulation complète après *abstraction*. Un parallèle peut être fait avec la classification de Smoliar et al. (1996) entre les méthodes *expressives* et *sémantiques* pour la recherche d'information.
3. **La dimension**
   Le résumé *mono-document* sera plutôt *indicatif* alors que résumé *multi-document* essaie d'éviter un examen de tous les documents.
4. **Le contexte**
   Le résumé peut être *générique* ou prendre en compte le besoin de l'utilisateur au travers d'une *requête*.

Un résumé automatique peut prendre en compte chacun des paramètres précédents au travers du besoin de l'utilisateur, mais dans la plupart des approches, ils sont fixés en fonction de l'application. Cependant, le cas du résumé de parole offre plus de possibilités car le média de destination peut être textuel ou parlé, impliquant une synthèse vocale et/ou la participation des locuteurs d'origine.

2.2.1 Évaluation

Nous allons présenter l'évaluation du résumé automatique sous le regard de la campagne *Document Understanding Conference* (DUC) organisée par NIST[9]. Les tâches liées à cette évaluation sont décrites plus en détail dans le chapitre 6. La tâche principale est de résumer une vingtaine de documents en fonction d'un besoin utilisateur avec une longueur maximale de 250 mots. Comme pour la plupart des tâches simulant une part de compréhension, l'évaluation d'un résumé est difficile. En effet, deux personne ne produisent pas le même résumé sur des mêmes contraintes de données, de besoin, de temps : il est difficile de définir le résumé parfait que doit générer un système. En effet, un résumé de même qualité peut être produit en utilisant des mots différents mais de même sens ; ou au contraire il est possible d'utiliser les mêmes mots tout en changeant leur ordre, ou en introduisant des négations, afin de détourner le sens. Les figures 2.3, 2.4 et 2.5 présentent un *topic* DUC, deux résumés et un document associés à ce *topic*. La table 2.2 montre les propriétés de quelques méthodes d'évaluation du résumé automatique.

Il existe des évaluations de sous-objectifs, très dépendantes de l'approche implémentée. Ces méthodes sont généralement peu représentatives du résultat final, mais facilitent l'interprétation des performances du système. Le résumé par extraction, par exemple, nécessite un module de sélection des phrases importantes dans les documents. Cette sous-tâche peut être évaluée en comparant la sélection d'un système auto-

[9]Site web : `http://doc.nist.gov`, visité en décembre 2006. DUC est organisée chaque année depuis 2001 et au moins jusqu'en 2007. Une autre évaluation, SUMMAC (Mani et al., 2002), est moins récente.

Tâche	Obj.	Rep.	Op.	Auto.	Type
Sélection de phrase	sous	- -	- -	+	intrinsèque
Compression de phrase	sous	- -	- -	+	
Tâches d'analyse linguistique	sous	- -	- -	+	
Questions linguistiques	semi	+	-	-	intrinsèque
Rouge	quasi	-	+ +	+	
Pyramids	quasi	+	+	-	
Besoin utilisateur (*responsiveness*)	pseudo	+ +	+ +	-	extrinsèque
Application	réel	+ + +	+ + +	-	

TAB. 2.2: *Différentes évaluations du résumé automatique et leurs caractéristiques principales telles que le réalisme de l'objectif (Obj.), la représentativité de l'évaluation (Rep.), l'opacité par rapport aux performances des composantes du système sous-jacent (Op.), la possibilité d'automatiser l'évaluation (Auto.), et le type d'évaluation (Type). Cette classification est inspirée d'un travail de groupe lors de l'atelier DUC 2005.*

matique avec celle d'un opérateur humain. Le problème provient de l'observation que deux phrases mises ensemble n'apportent pas la même information que lorsqu'elles sont séparées. La compression de phrase correspond à éliminer l'information inutile afin de gagner de la place dans un résumé de taille fixe et y introduire plus d'informations (Hori et al., 2003a). Cette sous-tâche reste difficile à évaluer comme pour le résumé automatique car même si aucune reformulation n'est autorisée, différents évaluateurs ont tendance à produire différentes formes compressées de référence. Plus généralement, l'ensemble des tâches d'analyse linguistique — comme la résolution des références ou la détection de l'argumentation — peuvent être évaluées comme un sous-objectif. De bonnes performances sur l'ensemble de ces tâches n'induisent pas forcément la qualité *in fine* du résumé.

Topic : D0604D
Title : anticipation of and reaction to the premier of Star Wars Episode I – The Phantom Menace
Narr : How did fans, media, the marketplace, and critics prepare for and react to the movie ? Include preparations and reactions outside the United States.

TAB. 2.3: *Exemple de topic (sujet) DUC 2006 sur la sortie du film « Star Wars : la Menace Fantôme » (D0604). Le champ Topic indique l'identifiant du topic ; le champ Title représente le titre et le champ Narr explicite les sous-thèmes que le résumé devra aborder.*

La campagne DUC évalue les résumés sur leur forme et leur fond. Pour la forme, le logiciel *Summarization Evaluation Environment*[10] (SEE) permet à des juges de noter manuellement la qualité des résumés selon les critères linguistiques suivants :

Q1 : la *grammaticalité* (problèmes de formatage, de capitalisation, d'omissions, qui rendent difficile la lecture du texte) ;
Q2 : la *redondance* (répétitions non nécessaires d'expressions ou de noms d'entités lorsqu'un pronom aurait suffi) ;

[10] http ://www.isi.edu/ cyl/SEE

Q3 : la *clarté* des références (difficulté à déterminer le référent d'un pronom, connecteurs logiques non satisfaits...) ;
Q4 : la *focalisation* (le résumé ne doit pas contenir d'informations hors de sa thématique) ;
Q5 : la *structure* et la *cohérence* (le résumé doit être organisé et non un amas d'informations).

Prior to its premiere on May 21, 1999, "Star Wars : Episode 1 : The Phantom Menace," the film's producer, Lucasfilm Ltd., and its distributor, 20th Century Fox, engaged in a massive campaign of pre-release hype. A 2-minute 10-second trailer for the film was publicized on the Lucasfilm Web site and screened in late 1998 in 26 states on 75 screens. It was shown on "Entertainment Tonight" and "Access Hollywood." Fans watched the trailer several times. Fans worldwide exchanged information about the film on the Internet. The press reported on buzz about the film. The Atlanta Journal-Constitution covered the trailer's debut and featured articles on the film, as did USA Today.

Lucasfilm Ltd. orchestrated pre-release efforts of fast food companies, toy manufactures, and video game manufactures As the film's release date approached, merchants geared up for what was expected to be one of the biggest rollouts of merchandise in history. Fans began frantic purchases of merchandise from stores and on-line. Media attention also intensified. George Lucas, the film's director appeared on two segments of "60 Minutes." Vanity Fair devoted cover stories to the film and its stars.

Fans lined up at theaters, often in "Star Wars" costumes, to see the film. This scene was repeated in other countries, as Japan, when the film was released. Initial box office sales indicated that receipts for the film could set a record. The reviews of the film were generally negative. Critics praised the special effects but felt that the film's characters were less than compelling.

In preparation for the premier of "Star Wars : Episode I - The Phantom Menace" in May 1999, a 2-minute 10-second trailer (or preview) was released in November 1998. "Star Wars" fans flocked to theaters throughout the United States, not to see the featured films, but to see the "Star Wars" trailer repeatedly. A German college student, not content to wait until the show's opening overseas in June, proposed a trip to the U.S. to see the movie in May and received eager responses from dozens of fans throughout Europe. Licensed merchandise from dolls to shampoo, featuring "Star Wars" characters and themes was marketed with anticipated sales of $1 billion. As May approached some fans stood in line as long as three weeks in advance to get tickets. Estimates of anticipated ticket sales were as high as $1.78 billion. Enthusiasm was shared by fans in Japan including 30 who flew to Los Angeles for the show's opening there.

The opening days were impressive. "The Phantom Menace" reached the $100 million benchmark in its first five days-a record. Crowds diminished after the first week and some reviews were less than completely laudatory. It had grossed $205 million by the 13th day, but by the 23rd day its take of $271.3 million placed it only 11th among cinema moneymakers. In England 13 million "Star Wars" books were printed to meet anticipated demand, but only 3 million were sold. The hype was over and the headline read "'Star Wars' Bombs in Britain."

TAB. 2.4: *Comparatif de deux résumés de référence DUC (provenant respectivement des experts J et E) pour le topic D0604 sur la sortie du film « Star Wars : la Menace Fantôme ». Alors que les thèmes abordés dans ces résumés sont similaires (la bande annonce, le merchandizing...), la structure générale est différente et les informations présentées ne se recouvrent pas complètement.*

Une note de 1 à 5 est donnée pour chacune de ces questions linguistiques, 1 correspondant à une très mauvaise qualité et 5 à l'équivalent d'une production humaine. Cette évaluation par un expérimentateur humain est très importante dans le cas d'un résumé textuel car elle reste à notre connaissance la seule façon de juger la forme d'un

résumé. Par contre, les critères sont dépendants du média dans lequel est formulé le résumé : l'audio ou la vidéo ne sont pas soumis aux même contraintes. Par exemple, nous proposons quelques critères plus précis qu'imposerait un résumé parlé :

Q6 : Le contenu vocal doit être agréable à écouter (rythme, clarté, environnement). Une voix de synthèse, par exemple, peut ne pas paraître naturelle ; un fond sonore trop présent dégrade la clarté du message ; une voix trop aiguë peut agacer l'auditeur.

Q7 : La prosodie doit être régulière et en accord avec le message. Il est plus difficile de comprendre une question ayant la prosodie d'une phrase déclarative. Une phrase coupée dont la prosodie n'a pas la forme attendue est désagréable à écouter.

Q8 : Si le résumé contient de la parole rapportée, celle-ci doit être différenciée de la narration. Sa présence doit être justifiée en montrant par exemple une prise de position.

Q9 : La locution du narrateur doit être claire (pas d'hésitations, de reprises, de réparations ...). Ces phénomènes sont souvent dus à un état émotionnel ou une pathologie du locuteur. Le narrateur doit montrer un état neutre. Cette neutralité permet de le différencier des intervenants en condition spontanée.

Q10 : L'identité des intervenants doit être explicitée.

Q11 : Les événements sonores hors parole doivent être décrits (ou attendus) et avoir une pertinence vis-à-vis du message. Par exemple, l'auditeur s'attend à entendre des explosions lors de l'intervention d'un reporter de guerre, mais ne comprend pas une accroche musicale inattendue.

De tels critères ont une incidence certaine sur les méthodes utilisées pour la génération du résumé, comme par exemple la synthèse vocale ou l'identification du locuteur, tâches qui peuvent elles-mêmes être évaluées séparément. Il n'existe pas à notre connaissance d'évaluation du résumé parlé, mais les critères Q6 à Q11 consitituent une base intéressante pour en construire une.

La réponse au besoin est une évaluation de fond du résumé. DUC a choisi de l'évaluer selon des critères manuels et automatiques (afin de réduire les ressources nécessaires à l'évaluation de l'amélioration d'un système). Des juges évaluent le fond en se posant la question : « le résumé répond-t-il au besoin de l'utilisateur ? ». Ce besoin est exprimé dans DUC sous la forme d'un thème et de requêtes en langue naturelle, mais il peut être exprimé dans une forme totalement différente et impliquer des contraintes beaucoup plus complexes à évaluer. DUC propose aussi une évaluation manuelle qui implique le fond et la forme. Les juges doivent répondre à la question « combien d'argent auriez-vous donné pour ce résumé ? ». Ces deux questions donnent lieu à une note entre 1 et 5 et s'appellent dans DUC, *content responsiveness* et *overall content quality*.

DUC propose ensuite une évaluation du fond moins opaque que la note générale de *responsiveness*, sous le nom de *Pyramids* (Nenkova et Passonneau, 2004). Cette évaluation manuelle demande le découpage des résumés de références en informations élémentaires (*Summary Content Units*, SCU ; « Bush a été battu par les démocrates » et « Les démocrates ont gagné la chambre des représentants et le sénat » seront considé-

<DOC>
<DOCNO> APW20000124.0232 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 2000-01-24 22:27 </DATE_TIME>
<BODY>
<HEADLINE> 'Star Wars' Bombs in Britain </HEADLINE>
<TEXT>
<P>
LONDON (AP) – A failed bet on a "Star Wars" craze in Britain has left a publisher with 10 million unsold books and a vacancy for the post of chief executive.
</P>
<P>
British children's books publisher Dorling Kindersley announced Monday that its chief executive has resigned because of "a seriously misjudged" overinvestment in books tied to the launch of the latest movie in the saga, "Star Wars : Episode I, The Phantom Menace."
</P>
<P>
Expecting a Christmas rush, James Middlehurst arranged for the group to print 13 million of the books in the 18 months to Dec. 31 last year. Sales totaled 3 million.
</P>
<P>
Company chairman Peter Kindersley blamed the "Star Wars" debacle for more than half of a pretax loss of $41 million which it expects to announce in the spring.
</P>
</TEXT>
</BODY>
<TRAILER> AP-NY-01-24-00 2227 </TRAILER>
</DOC>

TAB. 2.5: *Exemple de document DUC pour le topic D0604 sur la sortie du film « Star Wars : la Menace Fantôme ». Le document illustre un aspect traité dans le résumé de l'expert E.*

rées comme équivalentes), dont une pondération est donnée par le nombre de résumés qui les contiennent. Puis, ces informations élémentaires sont mises en relation manuellement avec celles contenues dans un résumé automatique, en autorisant des imprécisions, des généralisations et des spécialisations. Le score du résumé est donné par le nombre d'unités en commun, sachant que les unités de poids supérieur (les plus importantes) doivent apparaître avant d'obtenir la validation des unités de poids inférieur.

Lin (2004) propose une évaluation automatique fortement corrélée avec l'évaluation humaine à travers *Recall-Oriented Understudy for Gistering Evaluation* (Rouge). Cette mesure fait intervenir la différence entre la distribution des mots d'un résumé candidat et celle d'un ensemble de résumés de référence. Copeck et Szpakowicz (2004) estiment qu'il faut 30 résumés pour construire un résumé moyen de référence représentatif ; DUC n'en produit que 4 pour l'évaluation (pour des raisons de coût), n'ayant pas observé de différence fondamentale dans les mesures Rouge lorsque plus de résumés sont disponibles. Fortement utilisé durant les campagnes DUC pour lesquelles elle représente l'évaluation automatique du fond, ce genre de mesure est en constante amélioration et tend à intégrer des éléments conceptuels (notion de *Basic Elements* : entités-

relations). Il s'agit de se rapprocher d'un espace de représentation de l'information et de s'éloigner de l'espace d'instanciation linguistique. Une mesure automatique est indispensable à l'amélioration des systèmes car elle prend beaucoup moins de temps et ne demande pas d'opérateur humain[11]. Rouge-2 et Rouge-SU-4 se sont imposées dans DUC pour l'évaluation automatique. Rouge-2 correspond au nombre de bigrammes en commun entre un résumé automatique et l'ensemble des résumés écrits à la main (R_n avec $n = 2$ dans l'équation 2.10). Rouge-SU-4 correspond au rappel en « bigrammes à trous » (*skip units*) de taille maximum 4 (RSU_n avec $n = 4$ dans l'équation 2.11). La table 2.6 regroupe quelques exemples d'éléments utilisés dans Rouge.

$$R_n(hyp, ref) = \frac{|\{w_i, \cdots, w_{i+n-1}\}_{ref} \cap \{w_i, \cdots, w_{i+n-1}\}_{hyp}|}{|\{w_i, \cdots, w_{i+n-1}\}_{ref}|} \tag{2.10}$$

$$RSU_n(hyp, ref) = \frac{|\{w_i, w_{j<i+n+1}\}_{ref} \cap \{w_i, w_{j<i+n+1}\}_{hyp}|}{|\{w_i, w_{j<i+n+1}\}_{ref}|} \tag{2.11}$$

Phrase	**le chat boit du lait**
2-gram	le-chat, chat-boit, boit-du, du-lait
2-gram (lemmes)	chat-boire, boire-lait
SU-2	*2-grams* + le-boit, le-du, chat-du, chat-lait, boit-lait
BE	chat-le-dét., boire-chat-sujet, lait-du-prép., boire-lait-comp.

TAB. 2.6: *Illustration de différents découpages pour le calcul de Rouge. Les éléments de base sont : les n-grammes (lemmatisés ou non), les bigrammes à trous (Skip Units, SU) et les triplets d'entités-relations (Basic Elements, BE).*

Des pré-traitements peuvent être appliqués, comme la suppression des mots-outils (qui représentent en général 50% des occurrences des unigrammes), ou l'utilisation de dictionnaires de synonymes pour améliorer la corrélation avec les résumés écrits à la main. Un des inconvénients de Rouge est qu'un bon choix des sous-séquences de mots autorise la génération de résumés ayant un meilleur score que les résumés écrits manuellement au détriment de la forme. Notamment, les méthodes à base d'apprentissage maximisant le critère Rouge doivent absolument contrebalancer cet effet par des contraintes linguistiques.

Les méthodes proposées pour évaluer la qualité d'un résumé requièrent soit un jugement sur le moment (évaluation manuelle), soit la création de références *a priori* (évaluation automatique). Dans les deux cas, il est observé que l'accord entre les juges est généralement faible (Minel, 2004) et qu'il faut multiplier les jugements pour obtenir des résultats significatifs. La mesure Kappa (Fleiss, 1971) représente le degré d'accord entre les juges $P(A)$ au delà du hasard $P(C)$ (équation 2.12).

$$\kappa = \frac{P(A) - P(C)}{1 - P(C)} \tag{2.12}$$

[11] Les organisateurs de DUC estiment à 3000 heures le temps passé pour réunir les documents, créer les références et évaluer les soumissions (30 participants et 50 *topics*).

La même problématique se retrouve dans les mesures automatiques comme Rouge, pour lesquelles une méthode de *Jacknife* améliore la robustesse. Lorsqu'un résumé est évalué par rapport à N résumés de référence, l'évaluation est répétée N fois en retirant à chaque fois un résumé de référence. Le score final est défini par la moyenne des scores de chaque itération. Cette approche permet de calculer des intervalles de significativité par analyse de la variance ANOVA (Lindman et al., 1976).

Après ce tour d'horizon de l'évaluation du résumé, nous allons voir les approches les plus populaires pour le résumé textuel et les contraintes spécifique à l'utilisation d'un média parlé dans ce domaine.

2.2.2 Résumé par extraction

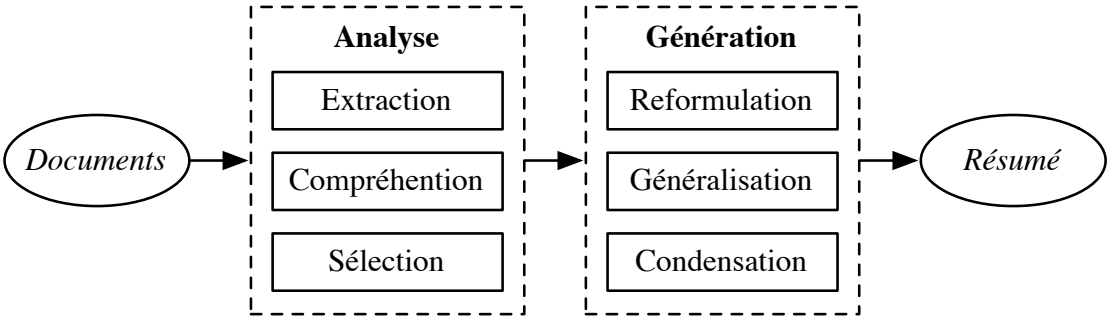


FIG. 2.5: *Processus de création d'un résumé, en deux étapes : une étape d'analyse de l'information contenue dans les documents d'origine, suivie d'une étape de génération dans le média cible.*

Le texte véhicule des idées, des concepts et des informations. Le résumé automatique de texte nécessite avant tout de dégager les informations importantes et de les resynthétiser avec un ratio de réduction fixé. Il a été observé qu'environ 70% des phrases utilisées dans des résumés réalisés par des opérateurs humains étaient empruntées au texte d'origine sans modification (Lin et Hovy, 2003; Jing, 2002). Il est dans ce cas tout à fait envisageable de rechercher avant tout ces phrases transposables sans nécessiter de «compréhension/synthèse» complète. Nous avons déjà vu qu'il existe deux grandes approches pour le résumé automatique. L'approche par extraction vise à trouver parmi les phrases d'origine, les plus susceptibles d'être réutilisées, en partie ou en totalité, dans le résumé généré. Ce genre d'approche implique une étape de sélection des phrases suivie d'une étape de placement des phrases dans le résumé et d'amélioration de la forme. L'alternative est d'identifier les informations importantes du domaine et de les mettre en correspondance avec des patrons de résumé permettant de générer des phrases et les assembler en un discours cohérent. La première méthode est la plus appropriée pour des résumés indépendants du domaine car elle utilise généralement des statistiques alors que la seconde demande des connaissances du domaine.

Analyse

Les méthodes d'analyse de l'information comprennent souvent une phase d'extraction d'unités sémantiques suivie d'une phase de réduction du contenu à l'essentiel. Cependant, il existe des approches focalisées sur le contenu sémantique et ignorant la forme de surface des phrases. Ou, au contraire, indépendantes du contenu mais utilisant des caractéristiques de surface indicatrices de l'importance d'une phrase. Toutefois, l'utilisation conjointe de ces critères est la plus répandue.

Les traitements linguistiques permettant l'extraction d'unités sémantiques à partir des mots sont généralement les mêmes que ceux appliqués en recherche documentaire avec pour spécificité que l'unité est la phrase au lieu d'être le document (voir section 2.1). Il n'est plus possible de compter sur l'effet quantitatif et redondant du document ; le résumé automatique est plus fortement demandeur de précision dans l'extraction des unités sémantiques et dans les divers pré-traitements, car les erreurs sont directement répercutées sur le résultat. Par exemple, Blair-Goldensohni et al. (2004) appliquent un étiquetage morpho-syntaxique des mots pour dégager des patrons de phrases et marquer les mots porteurs de sens en fonction de leur catégorie syntaxique. Les entités nommées sont des marqueurs de l'information du domaine utiles pour inférer la tendance d'une phrase à apparaître dans le résumé (Bergler et al., 2004). Dans un souci de précision de l'information présentée au sein d'une phrase, Vanderwende et al. (2004) et Witte et Bergler (2003) appliquent une résolution des anaphores pronominales et des groupes nominaux. Ces anaphores et ces entités nommées sont également utiles lors de la phase de génération du résumé pour améliorer la clarté des références. Le contenu de surface (les mots) peut être projeté dans un réseau sémantique tel Wordnet, comme le font Doran et al. (2004) pour tirer parti des relations entre les éléments (synonymie, généralisation...). Finalement, une analyse grammaticale complète aide à l'extraction de paires d'entités-relations représentatives du message (Hovy et al., 2005).

Les descripteurs sémantiques représentent chaque phrase sous la forme d'un point dans un « espace sémantique », permettant de sélectionner celles qui répondent le mieux au besoin de l'utilisateur. Cette sélection est réalisée de façon similaire à la tâche de recherche documentaire avec pour contrainte supplémentaire qu'il ne faut pas proposer plusieurs fois la même information. Pour cela, la sélection de phrases est formulée comme un problème d'optimisation consistant à maximiser la couverture de l'information présentée tout en minimisant sa redondance. Ce problème est en partie résolu par des algorithmes de classification non supervisée afin de partitionner l'espace en N classes et de sélectionner la phrase la plus représentative de chaque classe. Seki et al. (2004) implémentent un *clustering* hiérarchique ascendant et comparent diverses méthodes d'agrégation : le critère de Ward semble donner les résultats les plus équilibrés. Dans la même optique, *Maximal Marginal Relevance* (MMR) est un algorithme glouton qui sélectionne les phrases en fonction de leur similarité avec le besoin utilisateur et leur dissimilarité avec ce qui a déjà été sélectionné (Goldstein et al., 2000; Erkan et Radev, 2004; Mori et Sasaki, 2002). La formule de MMR est explicitée par l'équation 2.14, dans laquelle mmr_k correspond à la sélection de phrases à l'étape k, s_i au représentant dans l'espace sémantique de la phrase i, c au représentant du besoin utilisateur, $sim(a, b)$ à

un opérateur de similarité dans l'espace sémantique, et λ à un hyperparamètre fixé de manière empirique.

$$mmr_k = mmr_{k-1} \cup \{\hat{s}_k\} \tag{2.13}$$

$$\hat{s}_k = \underset{s_i \notin mmr_{k-1}}{\operatorname{argmax}} \left(\lambda sim(s_i, c) - (1-\lambda) \max_{s_j \in mmr_{k-1}} sim(s_i, s_j) \right) \tag{2.14}$$

Les implémentations classiques de MMR reposent sur le modèle vectoriel (VSM, voir section 2.1.4), en utilisant des vecteurs de mots et la similarité *cosine*. Au-delà de cette méthode rapide, le problème d'optimisation peut être résolu grâce à diverses techniques d'optimisation, comme le montrent (Alfonseca et al., 2004) en appliquant un algorithme génétique pour produire et sélectionner des résumés maximisant la couverture et minimisant la redondance. Gong et Liu (2001) ont une approche similaire à LSA, en remarquant que la décomposition en valeurs singulières effectue un partitionnement flou des phrases (voir section 2.1.4). La SVD effectue une décomposition de la matrice des occurrences de mots dans les phrases en trois matrices $U\Sigma V^T$. La matrice V donne une projection des phrases selon la base quasi orthogonale trouvée par la SVD. En comparant les axes de cette base à des thèmes latents, les phrases les plus représentatives de chaque thème (de projection maximale sur les axes principaux de la base) sont retenues pour construire le résumé.

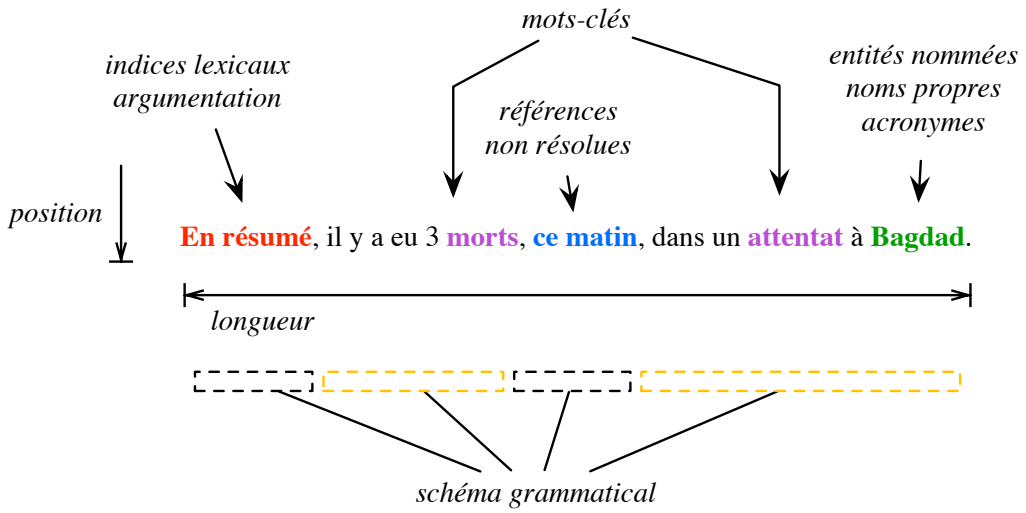


FIG. 2.6: *Quelques caractéristiques de la phrase ayant une influence sur sa sélection dans le résumé (position dans le document, longueur, indices lexicaux, références anaphoriques, mots-clés porteurs du contenu, entités spécifiques, schéma grammatical).*

L'information portée par une phrase n'est pas la seule caractéristique utile pour intégrer une phrase dans un résumé (Nobata et Sekine, 2004). La figure 2.6 illustre les caractéristiques pertinentes pour le résumé par extraction. Par exemple, il est observé que les phrases en début et en fin de document décrivent bien son contenu général (*position*). Les phrases courtes sont trop générales et les phrases longues sont trop spé-

cifiques (*longueur*). Il faut ajouter que les phrases qui contiennent des pronoms non résolus risquent de détériorer la cohérence finale du résumé (*références non résolues*). Des expressions permettent aussi de situer la phrase dans l'argumentation et de détecter les descriptions générales du contenu (*indices lexicaux*). Des mots-clés du domaine, ou tout simplement des mots-clés fréquents, ainsi que les mots morphologiquement indicateurs de contenu, comme les noms propres ou les acronymes, sont caractéristiques de phrases denses en informations (*mots-clés*, *entités nommées*, *acronymes*). Enfin, l'étude de la forme grammaticale de la phrase facilite la détection des propositions déclaratives les plus propices au résumé (*schéma grammatical*). Les caractéristiques utiles sont celles qui indiquent un fort contenu et qui permettent de maximiser la qualité linguistique du résumé produit. Les approches supervisées utilisent un corpus d'apprentissage (couples documents-résumés) pour résoudre le problème de classification consistant à sélectionner ou non une phrase pour le résumé. Par exemple, Kupiec et al. (1995) effectuent une classification bayesienne naïve et Daumé III et Marcu (2004) utilisent des machines à vecteur support (*Support Vector Machines*, SVM). Le problème est résolu de façon non supervisée par Torres-Moreno et al. (2002) grâce à une règle de décision prenant en compte une version normalisée de chaque statistique.

Des méthodes plus originales ont été proposées comme la création d'un graphe d'entités-actions (noms-verbes), sur lequel est appliqué un algorithme de popularité comme *PageRank* (Vanderwende et al., 2004). Il faut aussi noter l'approche par débruitage du texte après estimation du bruit généré par le canal qui le sépare du résumé (Daumé III et Marcu, 2001). Dans le cadre de l'évaluation DUC, Lacatusu et al. (2006) profitent des sous questions contenues dans l'expression du besoin utilisateur — à l'aide d'un système question-réponse — et appliquent une technique d'inférence textuelle (*textual entailment*) pour déterminer la redondance d'une phrase par rapport à une autre.

Génération

Le résumé par remplissage de patrons demande des modules de génération complexes, souvent à base de règles. Par exemple, Radev et McKeown (1998) effectuent une combinaison de patrons pour générer des transitions entre les phrases et utilisent des opérateurs pour construire l'argumentation. D'un autre côté, la génération la plus simple pour le résumé par extraction correspond à la juxtaposition des phrases sélectionnées jusqu'à l'obtention de la quantité d'informations voulue (ratio de réduction, nombre de phrases, nombres de mots). Les méthodes un peu plus avancées utilisent tout de même quelques post-traitements de surface pour améliorer la forme du résumé (Zajic et al., 2004) :

- compression des phrases par un modèle supprimant des sous-arbres syntaxiques ;
- remplacement des anaphores par les entités auxquelles elles se réfèrent (ou minimisation des références non résolues) ;
- normalisation des noms propres (leur première occurrence doit être le nom complet, les suivantes utiliseront uniquement le nom de famille...) ;
- suppression des annonces de discours rapporté (« *..., a dit le chef de la police* ») ;

– suppression des marques de l'argumentation générant des contre-sens lorsque les phrases sont utilisées hors contexte ;
– suppression du contenu entre parenthèses.

En plus de ces exemples de post-traitements, des efforts sont faits pour améliorer la structure et la cohérence des résumés. Daumé III et Marcu (2004) remarquent que l'ordre des phrases est important et qu'il peut être judicieux de sélectionner une phrase si elle est encadrée par deux phrases qui apparaîtront dans le résumé.

2.2.3 Spécificités de la parole

Un résumé automatique de parole est constitué à partir d'un flux audio parlé (entrées) et généré sous forme écrite ou parlée (sorties). La méthode la plus naturelle consiste à profiter des approches développées pour le texte, grâce à une étape de la transcription automatique du contenu parlé. Toutefois, dans ce cas, les traitements ne peuvent plus compter sur l'absolue précision de la forme écrite et doivent être capables d'intégrer les erreurs du processus automatique de transcription. Plusieurs questions doivent être explorées : à quel point les méthodes issues du texte sont-elles affectées par un contenu audio ? Est-il possible de les adapter à un contenu bruité ? Existe-il des différences fondamentales entre le texte et la parole qui permettent d'extraire des paramètres utiles supplémentaires ?

Si les recherches sur le résumé textuel se sont concentrées sur des contenus techniques et journalistiques[12], le résumé de parole est, pour l'instant, appliqué à trois sources de données : les émissions radiodiffusées, les dialogues téléphoniques et les réunions. Ces sources amènent néanmoins de nombreux challenges (McKeown et al., 2005). Christensen et al. (2003) montrent que les techniques classiques de résumé automatique sont portables aux journaux radiodiffusés et que les méthodes de sélection de phrases ont tendance à sélectionner des phrases dont le taux d'erreur de mots (provoqué par la transcription automatique) est plus faible que sur la globalité des données, un phénomène qui a aussi été observé par Murray et al. (2005). Cette différence en taux d'erreur s'explique par la relation entre l'importance d'une information et sa redondance dans les documents, selon la plupart des modèles de résumé automatique de textes.

Pour ce qui est de l'adaptation des méthodes de résumé et d'extraction d'informations, Zechner (2001) effectue une correction des malformations de l'élocution (*disfluencies*) comme les hésitations, les reprises, les coupures, les faux départs ou les pauses remplies. Il utilise des méthodes d'annotation de séquence sur le contenu linguistique et écarte les phrases ayant trop d'erreurs grâce à des mesures de confiance issues de la transcription. Il obtient ainsi une transcription propre, appropriée aux méthodes conçues pour le texte. Il reste malheureusement quelques erreurs qui réduisent les performances des techniques d'extraction de descripteurs sémantiques (entités nommées,

[12] Les évaluations DUC sont aujourd'hui concentrées sur les journaux, mais lors de l'atelier de la conférence en 2006 à New York, il a tout de même été discuté de migrer vers un contenu moins formel comme les *blogs*, et plus tard un contenu audio

relations, analyse grammaticale) et de nombreux travaux sont en cours dans ces domaines pour mieux coupler la transcription avec ces tâches intéressantes pour le résumé (Kubala et al., 1998; Van Noord et al., 2000).

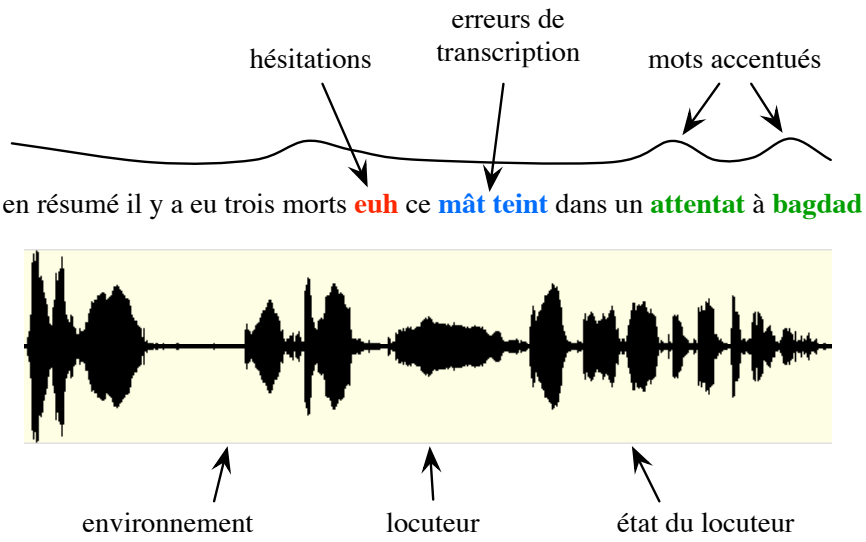


FIG. 2.7: *Paramètres supplémentaires de la parole par rapport au texte pour caractériser des phrases en vue de leur extraction pour créer un résumé (environnement acoustique, identité et type de locuteur, état du locuteur, hésitations, fiabilité de la transcription, et prosodie).*

La figure 2.7 donne quelques exemples des spécificités de la parole qui peuvent être ajoutées aux caractéristiques textuelles pour construire un résumé. Par exemple, Maskey et Hirschberg (2006) recherchent les paramètres permettant de se passer de la transcription pour faire un résumé. Plus précisément, ils tirent parti de la prosodie, et de la position des phrases dans le document à l'aide de Modèles de Markov Cachés (HMM). Bien que le modèle permette une amélioration significative par rapport au hasard, il ne s'applique que pour le résumé mono-document et les auteurs ne présentent pas de comparatif avec une méthode basée sur la transcription. Murray et al. (2005) utilisent un classifieur et des paramètres à la fois linguistiques, prosodiques et structurels pour générer des résumés de réunions et arrivent à des résultats intéressants en limitant la complexité des paramètres employés. Zhu et Penn (2005) comparent différentes approches pour cette tâche (MMR, une similarité sémantique, SVM et une régression logistique). Leurs conclusions sont que ces approches ont des performances très variables en fonction du degré de réduction et du taux d'erreur de mots de la transcription. La segmentation en phrases est un autre aspect important et non trivial du résumé par extraction. En effet, Mrozinski et al. (2006) comparent la segmentation de référence avec une segmentation automatique utilisant des caractéristiques linguistiques et une segmentation aléatoire. Ils observent que les performances en résumé baissent fortement lorsque la segmentation en phrases est de mauvaise qualité. Furui et al. (2004) de leur côté testent plusieurs unités de base pour l'extraction (mots, syntagmes « entre deux pauses remplies », phrases) dans le cadre d'un résumé parlé et observent que seules

les unités longues, comme la phrase, sont acceptables pour l'utilisateur. D'autres pistes sont explorées, comme l'intégration du rôle des locuteurs dans le résumé d'émissions radiophoniques (présentateur, reporter, invité, Barzilay et al., 2000) ou la cohérence des dialogues pour attacher les questions à leurs réponses (Zechner, 2002).

La plupart des méthodes appliquées à l'audio concentrent leurs efforts sur la partie analyse du processus de résumé. Seuls Hori et al. (2003b) essaient d'améliorer la partie génération en tirant parti des hypothèses de transcription pour réaliser une compression de phrase. Des modèles probabilistes d'informativité *a priori* des mots et de suppression de séquences de mots sont appliqués au graphe d'hypothèses de transcription dans le but de prendre en compte ces paramètres lors de la recherche de la meilleure hypothèse. Cette approche donne de bons résultats dans le cadre du sous-titrage d'emissions télévisées japonaises.

2.3 Conclusion

L'information parlée introduit de nombreux challenges comparé à la recherche d'information textuelle. La variabilité de la parole est le premier obstacle à l'extraction des descripteurs sémantiques d'un document. En outre, les nombreuses techniques d'analyse des phénomènes linguistiques sur le texte peuvent être appliquées à la transcription automatique du message parlé, au détriment d'erreurs proportionnelles au taux d'erreur de mots. Si l'impact de ces erreurs est relativement faible sur une tâche de recherche documentaire grâce à la redondance de l'information contenue dans les documents, il s'alourdit lors de l'extraction de descripteurs sémantiques de plus haut niveau et lorsque des structures plus petites que le document, comme la phrase, doivent être employées. Un premier challenge consiste à mieux coupler les tâches pré-transcription et post-transcription, en tirant parti des spécificités de la parole. Intégrer ces spécificités à la recherche d'information parlée représente un second verrou à lever. Toutefois, avant de s'intéresser à ces problèmes, il est indispensable de décrire les principes régissant la structuration de l'information parlée ; ceci est l'objectif du chapitre 3.