Intégration de contraintes d'interactivité dans le résumé

Sommaire

5.1 Portabilité à un média parlé de l'hypothèse d'extraction pour le résumé 97
5.2 Modèle général 100
5.3 Découplage fond-forme dans Maximal Marginal Relevance 102
5.3.1 Algorithme de sélection de phrases représentatives 102
5.3.2 Projection des phrases dans un espace pseudo-sémantique . . . 105
5.4 Conclusion 107

Ce chapitre expose une méthode de résumé par extraction d'un média parlé sous contraintes d'interactivité. La section 5.1 explore la validité d'une méthode de résumé par extraction sur des données radiophoniques. Une borne supérieure de performances Rouge est comparée aux performances d'approches triviales pour le résumé. Puis la section 5.2 définit un modèle de résumé par extraction séparant les traitements dépendant du besoin de l'utilisateur de ceux qui n'en dépendent pas. Les traitements coûteux doivent être indépendants de ce besoin pour satisfaire les contraintes d'interactivité. Le modèle proposé est intégré à *Maximal Marginal Relevance* (MMR) dans la section 5.3. Cette intégration est complétée d'une projection des phrases dans un espace pseudo-sémantique à l'aide de *Latent Semantic Analysis* (LSA). Nous concevons ainsi un système de résumé état de l'art prenant en compte les nouvelles contraintes liées à l'application visée.

5.1 Portabilité à un média parlé de l'hypothèse d'extraction pour le résumé

L'approche du résumé automatique de texte par extraction provient d'observations comme celles de Lin et Hovy (2003) et Jing (2002) qu'environ 70% du contenu d'un pa-

nel de résumés textuels écrits à la main (mettant en jeu entre 15 et 50 résumés) est extrait directement depuis les textes d'origine (copie de morceaux de phrases). Nous prenons pour hypothèse que cette observation, réalisée sur un corpus de résumés textuels, est transposable au résumé automatique de parole.

Afin de vérifier cette hypothèse, nous proposons une expérience de résumé automatique fondée sur les données de la campagne ESTER. Ces données ne contiennent pas de références pour l'évaluation du résumé. Par contre, elles sont segmentées manuellement en sections thématiques, dont les étiquettes sont à la fois représentatives du contenu sémantique et du contenu structurel. L'étiquetage permet de différencier les titres du journal radio-diffusé de son contenu[1]. Il est possible de tester grâce à ces données le succès d'une méthode de résumé par extraction pour générer les titres du journal (résumé cible) à partir du corps de celui-ci (documents source).

Rouge (Lin, 2004) est le critère utilisé pour valider la qualité des résumés (titres du journal) générés (voir section 2.2.1). Ce critère est connu comme étant fortement corrélé avec la qualité informative des résumés évaluée par des juges humains. Au cours de cette expérience, les résumés de trois systèmes fictifs de résumé sont comparés aux titres du journal :

1. Sélection au hasard des phrases ;
2. Sélection des N-premières phrases ;
3. Sélection de phrases maximisant le score Rouge.

Chacun de ces systèmes utilise les mêmes phrases du corps du journal, transcrites et segmentées manuellement (données de référence ESTER). Le taux de compression choisi correspond au rapport entre la longueur, en mots, du corps du journal et celle des titres visés[2]. Les deux premiers systèmes sont des systèmes triviaux couramment utilisés dans les évaluations du résumé par extraction. Ces mêmes évaluations ont prouvé que l'approche consistant en l'extraction des N premières phrases est difficile à battre dans le cadre d'un résumé mono-document (Over et Yen, 2003). Elle est beaucoup plus facile à surpasser dans le cas de résumés multi-documents. La sélection de phrases maximisant Rouge est obtenue par optimisation gloutonne en choisissant les N phrases de plus grand score individuel Rouge. Cette maximisation est sous-optimale, mais toute autre approche ne serait plus triviale (trouver la vraie borne supérieure est un problème NP-complet). La figure 5.1 compare les sorties des trois systèmes aux titres du journal pour un exemple tiré des données.

Une meilleure corrélation entre Rouge et les titres des journaux est obtenue en supprimant les mots outils des résumés hypothèse et référence. Cette suppression n'est pas appliquée dans les campagnes d'évaluation car il est tentant pour un participant de maximiser le nombre de mots à contenu informatif en supprimant les mots outils et en

[1] L'annotation d'origine n'est pas de bonne qualité : environ 35% des données ont été écartées pour créer des paires (titres – corps du journal) propres. La segmentation en sections thématiques n'a pas été remise en cause ; les titres contiennent souvent des données hors-sujet comme les annonce d'émissions.

[2] Les titres font en moyenne 177 mots ; les corps des journaux font en moyenne 3245 mots ; soit un ratio de 6% environ.

(1) Titres du journal de référence :
bonsoir à tous les grands titres de cette édition : double attentat suicide à Moscou faisant 18 morts et une cinquantaine de blessés , la police tchétchène retenue , la piste tchétchène plutôt retenue . Golfe : 7 policiers irakiens ont été tués et 45 autres personnes ont été blessées dans une explosion à Ramadi . sport tennis : l'américaine Serena Williams a remporté aujourd'hui le titre du simple dame des cent dix-septième internationaux de Wimbledon de tennis . les développements de ces titres , bonsoir à tous !

(2) Borne supérieure Rouge :
et en tennis l'américaine Serena Williams a remporté ce samedi , le titre du simple dame des cent dix-septième internationaux de Wimbledon de tennis sur gazon , troisième levée du grand chelem en battant en finale sa soeur aînée Venus par 2 sets à 1 . en Russie , un double attentat suicide perpétré par des femmes kamikazes est attribué par les autorités russes aux rebelles indépendantistes a fait 18 morts dont les 2 kamikazes et une cinquantaine de blessés lors d'un concert de rock ce samedi , à Moscou . dans le Golfe 7 policiers irakiens et un journaliste britannique ont été tués dans 2 attaques ce samedi en Irak , au lendemain de la diffusion d'une bande sonore attribuée à Saddam Hussein qui appelle à la guerre sainte contre les troupes occupantes .

(3) N premières phrases :
en Russie , un double attentat suicide perpétré par des femmes kamikazes est attribué par les autorités russes aux rebelles indépendantistes a fait 18 morts dont les 2 kamikazes et une cinquantaine de blessés lors d'un concert de rock ce samedi , à Moscou . selon le dernier bilan fourni par les autorités . un bilan précédent faisant faisait état de 20 morts le ministre russe de l'intérieur Boris Gryzlov ayant indiqué que 16 personnes étaient mortes sur le coup sans compter les kamikazes .

(4) Sélection aléatoire de phrases :
cette rencontre portera essentiellement sur la question des détenus palestiniens qu' Israël pourrait libérer ainsi que sur la suite de la mise en oeuvre de la feuille de route . l' agence Anatolie a annoncé ce soir qu'une partie des militaires turques arrêtés avaient été libérés . en Turquie euh , explosion dans un dans une station euh ,(de) service .

FIG. 5.1: *Exemple de résumés d'un journal (20030705_2300_2310_RTM_ELDA). Les titres du journal sont utilisés comme référence (1). Les résumés générés par trois systèmes fictifs sont présentés : la sélection de phrases maximisant Rouge (2),* $R_1 = 0.45$, $R_2 = 0.32$ *; les N premières phrases du corps du journal (3),* $R_1 = 0.26$, $R_2 = 0.08$ *; et une sélection aléatoire de phrases (4)* $R_1 = 0.04$, $R_2 = 0.00$.

faisant l'impasse sur la forme. Dans le cadre de nos expériences, un seul résumé de référence est disponible ; les mots outils sont supprimés afin de limiter l'effet sur le score de la réduction du nombre d'expressions anaphoriques permettant le recouvrement entre des expressions sémantiquement proches.

Les résultats de cette expérience sont présentés dans la table 5.1. Il est intéressant de noter que 39% des mots informatifs des titres du journal ont été retrouvés par la méthode maximisant la mesure Rouge. Ce résultat peut paraître peu élevé, mais il faut noter que la maximisation de Rouge est sous-optimale et que les synonymes ne sont pas considérés par la mesure. De plus, les scores des autres systèmes triviaux sont significativement moins bons quelle que soit la portée de Rouge. Ces résultats démontrent qu'une méthode de résumé par extraction peut être utilisée avec des données parlées de façon similaire à ce qui est fait sur le texte. Ce résultat ne remplace pas une évaluation de méthodes de résumé automatique de parole par extraction par des juges humain. De plus, il faut garder à l'esprit les conclusions de Banko et Vanderwende (2004)

Système fictif	Rouge-1	Rouge-2
Hasard	0.12756	0.03588
└ *écart-type*	*0.00507*	*0.00342*
N-premières	0.19007	0.08516
Borne supérieure	0.38618	0.24212

TAB. 5.1: *Performance des systèmes fictifs pour générer les titres des journaux ESTER à partir du corps du journal. Le premier système fait une sélection aléatoire de phrases (Hasard), le second utilise les premières phrases du corps du journal (N-premières) et le troisième repose sur la sélection de phrases maximisant les mesures Rouge. La borne supérieure indique qu'une méthode de résumé par extraction peut être utilisée sur des données parlées de façon similaire à ce qui est fait sur le texte. La sélection aléatoire est moyennée sur 50 initialisations différentes du générateur pseudo-aléatoire.*

selon lesquelles les méthodes par extraction doivent être outrepassées pour obtenir des avancées significatives dans le domaine du résumé multi-document.

Nous allons maintenant proposer un modèle général pour le résumé intégrant les contraintes de l'interactivité et d'un média parlé.

5.2 Modèle général

Une approche idéale pour le résumé serait de générer tous les résumés possibles et de sélectionner, par rapport au besoin de l'utilisateur, celui qui maximise à la fois la qualité du fond (quantité d'information et non-redondance) et celle de la forme (linguistique et acoustique). Comme il a été vu dans la section 2.2, le résumé par extraction est une approximation prenant pour hypothèse une indépendance et une complétude des phrases pour générer un résumé. Selon la plupart des méthodes, les scores de qualité du résumé peuvent être calculés comme la somme des scores des phrases qui le composent. Hassel et Sjöbergh (2006), par exemple, construisent un résumé textuel par extraction optimal au sens de sa représentativité du contenu d'un document mais sans prendre en compte, ni la forme, ni la redondance du fond. Le nombre de résumés à générer pour obtenir une solution exacte à ce problème est un arrangement de p phrases parmi n phrases possibles, soit $n!/(n-p)!$ résumés différents (par exemple, pour $n = 100$ et $p = 10$, il y a un peu moins de 10^{20} résumés). Il faut donc trouver une solution approchée dont le coût en temps soit suffisamment raisonnable pour permettre une interaction avec l'utilisateur.

Dans le contexte d'une interaction avec un utilisateur, le temps nécessaire pour donner une réponse à l'utilisateur est un facteur de qualité primordial. Pour le résumé automatique, ce temps est dépendant de la quantité de données traitées et de la complexité des algorithmes mis en œuvre. Des approximations peuvent parfois être employées pour réduire le temps de traitement d'un algorithme trop complexe. Nous allons présenter un modèle général pour minimiser l'effet de ces approximations et améliorer le temps de réponse du système.

La figure 5.2 montre différentes classes de modèles possibles pour le résumé auto-

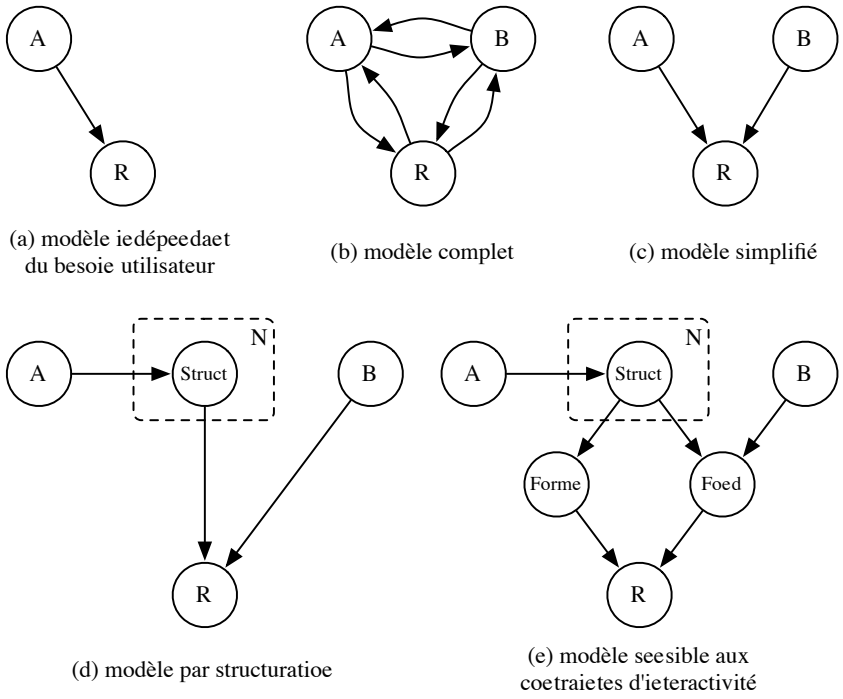

FIG. 5.2: *Différentes classes de modèles pour le résumé automatique. A représente l'acoustique, B le besoin utilisateur et R le résumé. Les modèles indépendants du besoin utilisateur déduisent le résumé de l'acoustique (a). Le besoin de l'utilisateur peut être intégré de différentes manières. Le modèle complet (b) prend en compte l'ensemble des dépendances possibles entres les composantes du problème. Le modèle simplifié (c) introduit une hypothèse d'indépendance entre l'acoustique et l'expression du besoin pour faciliter la résolution du problème (modèle classique). Le modèle par structuration (d) ajoute des sous-tâches de structuration pour faciliter l'émergence d'une sémantique dans l'acoustique. Le modèle sensible aux contraintes d'interactivité (e) introduit une séparation entre les paramètres liés à la forme du résumé, indépendants du besoin et ceux liés au fond du résumé, dépendants du besoin. La représentation proposée dans cette figure repose sur le formalisme des modèle graphique : une flèche représente une dépendance conditionnelle.*

matique. Dans cette figure, *A* représente l'acoustique, *R* représente un résumé audio satisfaisant l'utilisateur et *B* représente le besoin de l'utilisateur. Un premier modèle reflète les méthodes ne prenant pas en compte le besoin de l'utilisateur : le résumé est construit uniquement à partir des propriétés intrinsèques de l'acoustique. Pour introduire le besoin dans ce modèle, un modèle complet contenant toutes les interactions possibles entre les trois événements peut être imaginé. Ce modèle fait intervenir des aspects théoriques encore peu explorés, comme une dépendance possible des données et du besoin de l'utilisateur sur le résumé. Bien que très intéressant, le modèle complet n'est malheureusement pas réaliste à l'heure actuelle, surtout pour une application à ressources limitées. La plupart des approches en résumé automatique dépendant du besoin utilisateur sont fondées sur une hypothèse d'indépendance entre l'acoustique et le besoin. Cette hypothèse est illustrée par le modèle simplifié, ou modèle classique en résumé automatique (le résumé dépend de l'acoustique et du besoin). Cette indé-

pendance est discutable mais rend le problème beaucoup plus abordable. Ce modèle implique toutefois la constitution du résumé directement à partir des propriétés de l'acoustique, qui représente une quantité de données trop importante et pas suffisamment riche pour être exploitée correctement. Comment un besoin exprimé en langue naturelle peut-il être confronté directement à des données audio ? Un modèle par structuration emploie des couches de structuration du contenu acoustique avant d'aboutir au résumé. Grâce à ce type de structuration, une sémantique est extraite de l'acoustique et en annote le contenu. Différents éléments de cette structuration ont été étudiés au chapitre 3. Ce type de modélisation représente l'approche la plus intuitive et la plus répandue pour traiter des données audio. Pour rendre ce modèle sensible aux contraintes de l'interactivité, les descripteurs issus de la structuration sont séparés en deux classes, ceux qui auront un impact lié au besoin, et ceux qui peuvent être traités de façon indépendante du besoin. La figure 5.2 matérialise cette idée par la séparation des descripteurs liés à la forme de ceux liés au fond car les premiers sont naturellement moins influencés par le besoin. Néanmoins, cette séparation n'est pas nécessairement limitée à ces deux classes de descripteurs. L'approche par séparation limite l'approximation due aux contraintes d'interactivité à la seule composante liée au besoin. Le reste des traitements peut être mené à l'avance en utilisant des méthodes plus complexes. Pour le résumé par extraction, ceci se traduit par un degré de prédisposition des phrases à l'apparition dans un résumé.

La prochaine section est dédiée à l'intégration du modèle sensible aux contraintes d'interaction dans la méthode de résumé par extraction *Maximal Marginal Releveance* (MMR).

5.3 Découplage fond-forme dans Maximal Marginal Relevance

Cette section est organisée en deux sous-parties. Tout d'abord, la formulation de *Maximal Marginal Relevance* est subdivisée en paramètres dépendants du besoin de l'utilisateur (le fond) et de ceux qui peuvent être calculés par avance (la forme). Ensuite, une projection des phrases dans un espace pseudo-sémantique est proposée pour modéliser le fond.

5.3.1 Algorithme de sélection de phrases représentatives

La méthode générale de sélection de phrases pour le résumé par extraction est de partitionner l'espace informatif en groupes de phrases sur un thème, ou un événement important, puis de sélectionner une phrase représentative par groupe. Le résumé sera constitué de la juxtaposition des phrases représentant chaque groupe. Des contraintes d'interaction avec l'utilisateur dirigent néanmoins le choix vers une méthode hybride réalisant le partitionnement de l'espace informatif en même temps que la sélection des phrases représentatives. Cette méthode, dénommée *Maximal Marginal Relevance* (MMR), a été proposée par Goldstein et al. (2000) pour déterminer la sélection de phrases candidates dans un résumé par extraction.

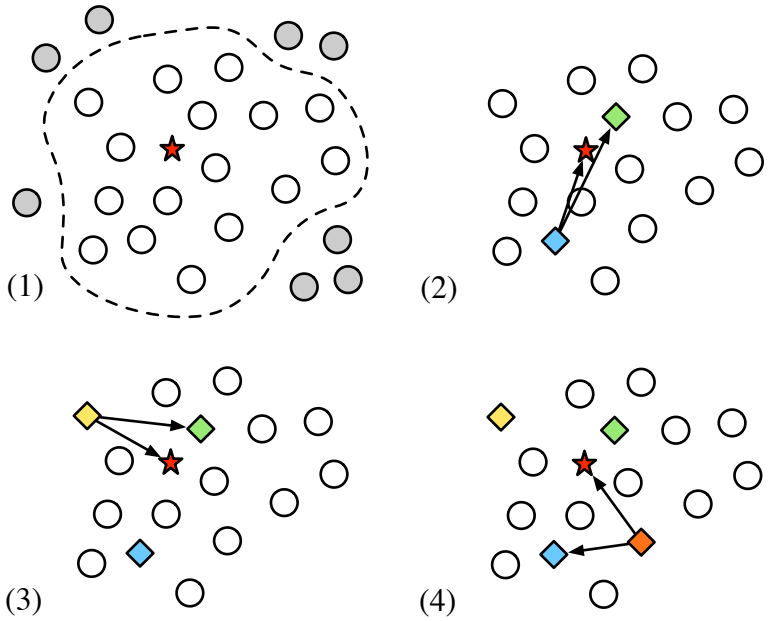

FIG. 5.3: *Illustration du fonctionnement de Maximal Marginal Relevance (MMR). La projection du besoin utilisateur est représentée par une étoile, les phrases candidates par des cercles et les phrases sélectionnées par des losanges. La première étape est d'écarter les phrases non pertinentes à l'aide par exemple d'une approche de recherche documentaire (1). La première phrase sélectionnée est celle qui est la plus proche du besoin. Puis, les phrases sont sélectionnées itérativement en fonction de leur distance à la projection du besoin, contrebalancée par leur redondance estimée par leur distance à la phrase déjà sélectionnée la plus proche (2,3 et 4).*

Dans les formulations suivantes, un résumé est constitué à partir d'un ensemble de documents thématiquement cohérents et d'un besoin utilisateur. Les documents sont découpés en phrases qui représentent la matière première de l'algorithme de résumé. À ce niveau, les phrases sont considérées comme indépendantes les unes des autres : leur origine et leur ordre sont ignorés. Une phrase est dénotée (s_i) pour la différencier d'une autre phrase (s_j). Pour l'instant, aucune hypothèse n'est faite sur le contenu des phrases, il est juste important de pouvoir les comparer deux à deux. Le besoin utilisateur est noté (b). La seule hypothèse nécessaire réside dans la possibilité d'évaluer une phrase en fonction de sa réponse au besoin.

MMR est une approximation gloutonne de la résolution du problème d'optimisation consistant à maximiser l'information tout en minimisant la redondance de l'ensemble de phrases sélectionnées pour le résumé. À chaque itération, l'algorithme détermine la phrase ($\hat{s}_k$) la plus proche de l'expression du besoin utilisateur (b) tout en étant la plus éloignée des phrases (s_j) sélectionnées auparavant. Cette phrase est ajoutée à la sélection et l'algorithme s'arrête lorsqu'une condition est remplie comme par exemple lorsqu'un nombre de phrases K, un nombre de mots ou un ratio de compression est atteint. La figure 5.3 illustre ce fonctionnement. Si n est le nombre de phrases à l'origine,

une implémentation efficace de MMR aura une complexité[3] en $O(n^2)$. L'équation 5.1 illustre la fonction de sélection d'une phrase à l'étape k.

$$\begin{aligned} mmr_0 &= \emptyset \\ mmr_k &= mmr_{k-1} \cup \{\hat{s}_k\} \\ |mmr_k| &< K \\ \hat{s}_k &= \underset{s_i \notin mmr_{k-1}}{\operatorname{argmax}} \left(\lambda sim_1(s_i, b) - (1-\lambda) \max_{s_j \in mmr_{k-1}} sim_2(s_i, s_j) \right) \end{aligned} \tag{5.1}$$

Dans la formulation originelle de MMR, $sim_1(\cdot)$ et $sim_2(\cdot)$ sont la similarité $cosine(\cdot)$ qui a fait ses preuves en recherche documentaire (voir section 2.1.4). Cependant, n'importe quelle similarité entre phrases est adaptée à ce problème. λ est un hyperparamètre[4] devant être ajusté empiriquement en fonction du cadre d'utilisation. Le modèle sensible aux contraintes des interactions utilisateur est introduit en modifiant la façon dont est sélectionnée la meilleure phrase à l'étape k (équation 5.2).

$$\hat{s_k} = \underset{s_i \notin mmr_{k-1}}{\operatorname{argmax}} \left(\lambda_1 \phi(s_i) + \lambda_2 \psi(s_i, b) - \lambda_3 \max_{s_j \in mmr_{k-1}} sim(s_i, s_j) \right) \tag{5.2}$$

Dans l'équation 5.2, $\psi(s, b)$ est le potentiel d'une phrase (s) pour le résumé en fonction de paramètres dépendants du besoin (b) ; $\phi(s)$ est le potentiel d'une phrase indépendamment du besoin. La limitation de la redondance du résumé est inchangée. Si $\phi(s)$ peut être précalculée et comporter des composantes gourmandes en ressources, $\psi(s, b)$ au contraire doit garantir un résultat rapide pour satisfaire les contraintes de l'interaction. La formulation classique de MMR peut être retrouvée en fixant $\phi(s) = 0$ et $\psi(s, b) = sim(s, b)$.

La fonction $\phi(\cdot)$ représente l'intérêt *a priori* pour une phrase dans le processus de sélection en fonction de ses caractéristiques indépendantes du besoin utilisateur. Le choix de ces composantes est difficile car il s'agit d'un compromis entre le temps de réponse du système et le potentiel d'intégration du besoin utilisateur. En réalité, il existe beaucoup de paramètres pour lesquels un lien direct au besoin utilisateur n'apparaît pas comme primordial. Par exemple, donner la possibilité à l'utilisateur de spécifier la longueur de phrase moyenne du résumé n'est pas indispensable. Il serait plus intéressant d'inférer ces paramètres directement de l'expression du besoin en langue naturelle, qu'ils soient explicités par le discours, ou latents. Les paramètres à prendre en compte sont de trois types (ces paramètres ont été illustrés par les figures 2.6 et 2.7 de la section 2.2) :

[3] Plus précisément, la complexité est de $n \times K$ calculs de la similarité entre deux phrase ou entre une phrase et le besoin, et $3 \times n$ en quantité de mémoire. Ces complexités peuvent être aussi bornées en ignorant de façon arbitraire les phrases répondant le moins au besoin de l'utilisateur.

[4] La nécessité de limiter la redondance dépend du contenu du résumé. Au lieu de faire varier λ en fonction du nombre d'itérations de l'algorithme (Murray et al., 2005), nous préférons normaliser la distribution des similarités $sim_1(\cdot)$ et $sim_2(\cdot)$ à chaque itération en leur imposant une moyenne nulle et une variance unitaire (standardisation des distributions).

1. les caractéristiques classiques de la phrase (longueur, position, anaphores non résolues...) ;
2. des mesures de confiance issues de la structuration (probabilité *a posteriori* de la transcription...) ;
3. les caractéristiques intrinsèques à la parole (prosodie, qualité d'élocution, identité du locuteur...).

Utilisées dans $\phi(\cdot)$, ces caractéristiques peuvent être impliquées dans n'importe quelle régression coûteuse mais performante, comme des SVM. Par contre, dans $\psi(\cdot)$, une méthode peu coûteuse comme une combinaison linéaire est indispensable. $\phi(\cdot)$ représente une décision optimale indépendante du besoin (ou dépendante d'un besoin utilisateur moyen), alors que $\psi(\cdot)$ intègre le besoin réel de l'utilisateur.

Dans le cadre de la chaîne de structuration Speeral décrite en section 3.1 et des compléments présentés dans le chapitre 4, nous pouvons introduire plusieurs éléments dans $\phi(\cdot)$. Pour la transcription, une confiance acoustique peut être calculée à partir du rapport entre la probabilité acoustique de la séquence de mots et la probabilité acoustique de la séquence de phonèmes non contrainte par les mots ; une confiance linguistique sera fonction du nombre de fois ou le système a utilisé un repli (séquence de mots non observée dans les données d'apprentissage) dans l'estimation de la probabilité linguistique. D'autres mesures de confiance sont présentées par Mauclair et al. (2006). La qualité de la segmentation en phrases peut être estimée par la probabilité marginalisée de sa frontière de début et de fin. La confiance linguistique dans les entités nommées extraites peut être calculée de la même manière que pour la transcription, mais sera moins bien estimée à cause de la faible quantité de données impliquées dans l'apprentissage des modèles n-grammes d'entités nommées. Une confiance plus représentative sera inférée à partir de la fréquence des entités retrouvées.

5.3.2 Projection des phrases dans un espace pseudo-sémantique

La similarité dans l'espace sémantique utilisée par MMR peut être fondée sur la plupart des méthodes de recherche d'information détaillées dans la section 2.1. Nous nous sommes concentrés sur $cosine(\cdot)$, mais afin d'outrepasser les limitations du modèle vectoriel classique (VSM), la base de l'espace sémantique est construite par *Latent Semantic Analysis* (LSA). Nous détaillons cette méthode dans les paragraphes suivants.

L'objectif de LSA est d'obtenir un espace modélisant les affinités contextuelles des mots comme approximation de leurs relations sémantiques. Contrairement à la formulation classique qui projette les requêtes (phrases) dans un espace réduit construit à partir des documents, nous suivons les travaux de Widdows et Peters (2003) et calculons une base unique sur un corpus de grande taille pour pouvoir projeter de nouveaux documents sans avoir à recalculer la base. Cette approche s'apparente au modèle vectoriel généralisé (GVSM). Ainsi, le vecteur représentant une phrase est la somme des vecteurs représentant les mots la composant dans l'espace LSA.

La construction de l'espace LSA demande tout d'abord la création d'une matrice

de cooccurrence entre les mots. Chaque ligne et chaque colonne de cette matrice représentent un mot et la cellule à l'intersection d'une ligne i et d'une colonne j contient le nombre de fois que les mots i et j se sont retrouvés ensemble dans un contexte donné (équation 5.3).

$$w_i^T \rightarrow \begin{matrix} \\ \\ \end{matrix} \begin{matrix} \quad w_j \\ \quad \downarrow \end{matrix} \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \tag{5.3}$$

La portée de la cooccurrence peut être la phrase, le document, ou un contexte de taille fixe. Par exemple, le tableau de contingence des bi-grammes est construit en limitant le contexte au mot précédent. Généralement, afin de s'affranchir de la nécessité de frontières fixes, une fenêtre glissante de n mots représente le contexte. Bien que la matrice de cooccurrences soit symétrique et creuse, elle demande quand-même une quantité de mémoire en $O(n^2)$, ce qui devient vite exorbitant lorsque le vocabulaire traité est grand. Pour réduire cet effet, il est courant de limiter le nombre de mots observés aux N mots les plus fréquents et de calculer leurs cooccurrences avec non pas l'ensemble du vocabulaire, mais une petite quantité de mots très fréquents et bien choisis, comme les entités nommées, afin de représenter des domaines sémantiques. Dans la suite, le rang de cette matrice va être réduit pour trois principales raisons :

1. la matrice est de trop grande taille pour assurer une faible complexité de calcul dans les traitements de grandes masses de données ;
2. l'estimation des cooccurrences n'est pas fiable par rapport à la distribution réelle (à cause de l'utilisation de synonymes, par exemple) ;
3. les erreurs de transcription bruitent cette matrice (dans le cas où des données textuelles ne sont pas disponibles pour en améliorer l'estimation).

La réduction du rang de la matrice se fait par décomposition en valeurs singulières (*Singular Value Decomposition*, SVD). Le principe est qu'une matrice réelle X peut être factorisée en 3 matrices U,Σ et V avec Σ une matrice diagonale positive, et U et V des matrices orthogonales (équation 5.4).

$$X = U\Sigma V^T \tag{5.4}$$

Σ contient les valeurs singulières σ_i de X, et U et V contiennent les vecteurs singuliers respectifs. Si les σ_i sont ordonnés de façon décroissante, le rang de la matrice Σ peut être réduit à k en annulant les valeurs singulières de rang supérieur à k. Cette réduction correspond à une approximation de X minimisant l'erreur au sens de moindres carrés (équation 5.5). La décomposition en valeurs singulières correspond à la minimisation de la corrélation entre les vecteurs singuliers. La base créée est souvent qualifiée de base « thématique » dans laquelle chaque vecteur représente un thème détecté automatiquement. Les mots sont exprimés dans cette base comme un vecteur de poids des

différents thèmes auxquels ils participent. Dans la pratique, il est difficile d'interpréter les thèmes détectés, mais l'espace créé représente bien les affinités lexicales des mots.

$$\hat{X} = U\Sigma_k V^T \tag{5.5}$$

La base de vecteurs singuliers U est utilisée pour exprimer un contexte d (requête, phrase, paragraphe, document...) en fonction des mots qui le composent. Dans l'équation 5.6, d est un vecteur sur l'ensemble du vocabulaire, fonction de la fréquence des mots dans le contexte observé et $\hat{d}$ est la projection de ce vecteur dans l'espace sémantique.

$$\begin{aligned} d &= (w_0, \cdots, w_n)^T \\ \hat{d} &= \Sigma_k^{-1} U^T d \end{aligned} \tag{5.6}$$

Cette approche a pour principal intérêt de représenter chaque unité informative par un vecteur dans un espace de dimension réduite. Ceci diminue le coût de calcul d'une similarité entre deux éléments. L'espace est considéré comme un modèle du monde (de tout ce qui peut y être instancié) et peut être généré à partir de données externes, disponibles en grande masse. Par contre, par rapport au modèle vectoriel simple utilisé en recherche documentaire, il n'est pas possible de créer un index inversé accélérant — par exemple — le calcul pour trouver l'élément le plus proche d'un élément donné (tous les éléments doivent être parcourus). Comme dans de nombreuses approches, la similarité *cosine* est utilisée pour comparer deux éléments ; ce n'est autre que le cosinus de l'angle entre les deux vecteurs représentant ces éléments :

$$\begin{aligned} cosine(a,b) &= \frac{a \cdot b}{|a||b|}, \\ &= \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}}. \end{aligned}$$

Suivant des travaux de Murray et al. (2005), l'espace sémantique ainsi créé est utilisé pour calculer la similarité entre deux phrases dans l'algorithme de résumé automatique présenté en 5.3.1, sous le nom de MMR-LSA.

5.4 Conclusion

Nous avons présenté dans cette section une intégration de contraintes de complexité dans une approche classique de résumé automatique (MMR). Celle-ci se traduit, dans le calcul de l'adéquation de phrase pour le résumé, par la séparation entre les caractéristiques provenant de la forme et celles provenant du fond. La forme est considérée comme indépendante du besoin alors que le fond reste lié à ce dernier. L'adéquation d'une phrase au besoin de l'utilisateur (sur le fond), est calculée après projection des

phrases et du besoin dans un espace pseudo-sémantique (LSA). Nous allons maintenant, dans le chapitre 6, évaluer le système de résumé automatique conçu à partir de ces approches sur des données textuelles, puis simuler sur cette même tâche le type d'erreurs imposé par une structuration automatique du contenu parlé.

Chapitre 6

Évaluation indirecte sur le texte

Sommaire

6.1 La campagne d'évaluation Document Understanding Conference . . 109
6.1.1 Descriptif de la soumission LIA-Thales 111
6.1.2 Résultats sur DUC 2006 118
6.2 Simulation de l'impact d'un contenu parlé 125
6.2.1 Cadre expérimental 125
6.2.2 Résultats sur les données dégradées 127
6.2.3 Interprétation des résultats 130
6.3 Conclusion 131

Il n'existe pas à notre connaissance de campagne d'évaluation du résumé parlé, même hors des conditions nous intéressant (résumé multi-document, sur le français, portant sur des données radio-diffusées et en association avec une tâche de recherche d'information). Il est tout de même possible d'évaluer indirectement la qualité de l'approche proposée en employant des données d'évaluation portant sur le texte et non sur la parole. Pour cela, une participation conjointe LIA-Thales à la campagne DUC 2006 a permis la validation de l'approche sur des données provenant de journaux, un type de données proche des émissions radio-diffusées (section 6.1). De plus, nous allons simuler l'impact de la structuration automatique d'un contenu parlé sur les données DUC pour avoir une idée de l'évolution des performances du système en conditions dégradées (section 6.2).

6.1 La campagne d'évaluation Document Understanding Conference

L'édition 2006 de DUC s'est concentrée sur la stabilisation de la campagne avec une tâche très similaire à l'année précédente afin de vérifier et consolider les acquis[1]. L'ob-

[1]Descriptif de la tâche disponible sur http://www-nlpir.nist.gov/projects/duc/duc2006/tasks.html, visité en janvier 2007.

jectif de l'évaluation est de générer un résumé de 250 mots ou moins (les résumés sont tronqués s'ils sont trop longs, il n'y a pas de bonus à faire plus court), à partir de 25 documents extraits de sources journalistiques et d'une description du besoin utilisateur. Le besoin utilisateur est formulé par un champ *titre* concis et un champ *description* qui liste le type d'information recherchée. En général, le champ *description* contient des sous-besoins du type :

- *Quels sont les causes, les conséquences, les problèmes de (...) ?*
- *Listez les types de (...). Quelles sont leurs particularités ?*
- *Détaillez chronologiquement, et/ou géographiquement, les événements liés à (...).*

Des illustrations de besoin utilisateur sont données dans la table 6.1. Bien que le besoin soit souvent formulé à l'aide de questions, ces dernières ne peuvent pas être traitées comme les questions fermées du type de celles apparaissant dans les tâches de questions-réponses. Ces questions n'ont clairement pas une réponse complète écrite dans un des document, mais elles mettent en jeu des capacités d'abstraction et de raisonnement.

D0617H – Le vol 990 d'EgyptAir
Qu'est-ce qui a causé le crash du vol 990 d'EgyptAir ? Détaillez les éléments de preuves, les théories et les spéculations.
D0629B – Les virus informatiques
Identifiez les virus informatiques ayant eu une propagation mondiale. Détaillez de quelle façon ils se répandent, les systèmes d'exploitation affectés, leurs pays d'origine, et leurs créateurs quand cela est possible.
D0641E – Le réchauffement climatique
Décrivez les théories concernant les causes et effets du réchauffement climatique et les arguments contre ces théories.

TAB. *6.1:* *Exemples de topics DUC 2006 traduits de l'anglais.*

DUC 2006 implique 50 *topics* (définitions de besoin utilisateur, requêtes, sujets ou thèmes) avec leurs 25 documents associés. Il faut noter que les documents fournis pour un *topic* sont pertinents et ne contiennent pas d'information hors-sujet. En terme d'évaluation, les résumés soumis sont notés manuellement sur le fond, la forme et une note globale prenant en compte à la fois fond et forme. Des évaluations automatiques Rouge-2, Rouge-SU4, et *Basic Elements* viennent les compléter grâce à des références produites par NIST, contenant 4 résumés par *topic*. Enfin, l'évaluation *Pyramids* (Nenkova et Passonneau, 2004) est produite par une partie des participants. La tâche telle qu'elle est décrite est très similaire à celle qui a été conduite en 2005. 34 participants ont soumis des résumés lors de l'édition 2006, 20 d'entre eux ont participé à l'évaluation par la méthode *Pyramids*.

6.1.1 Descriptif de la soumission LIA-Thales

Notre participation à DUC 2006 est exposée dans cette section ; les autres participants ont décrits leurs systèmes dans les actes de l'atelier de clôture de l'évaluation[2].

Principe

La tâche principale de DUC n'a pas évolué entre 2005 et 2006, ceci permet de profiter des données de 2005 pour affiner les paramètres d'un système destiné à l'évaluation en 2006. Cependant, le faible nombre de *topics* (50) impose une certaine prudence dans l'utilisation de ces données. En optimisant trop un système sur 2005, ce dernier risque de réduire ses performances sur l'évaluation 2006 à cause de la différence dans les thèmes traités et les résumés attendus. Cette nécessaire prudence nous conduit à développer une méthode minimisant le risque de sur-apprentissage et améliorant la robustesse de l'estimation des paramètres.

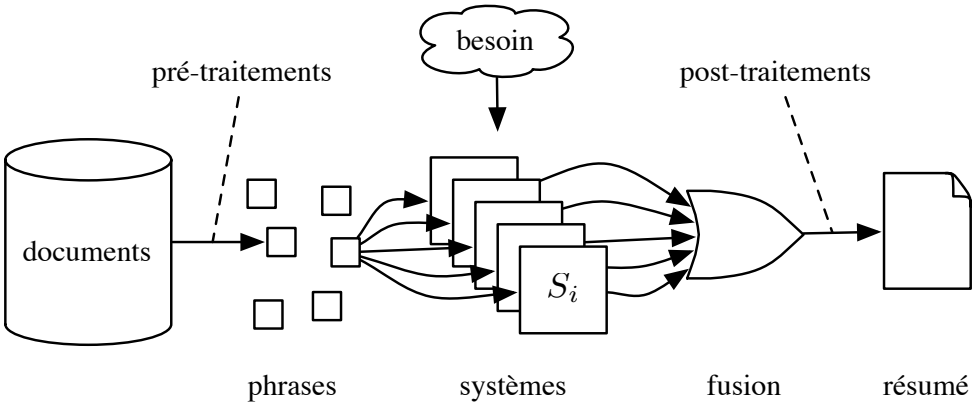

FIG. 6.1: *Schéma de fonctionnement du système LIA-Thales pour DUC 2006. Les documents sont découpés en phrases après application de pré-traitements. Puis différents systèmes établissent des listes de priorité d'inclusion des phrases dans le résumé. Les listes sont fusionnées pour trouver une sélection optimale de phrases. Des post-traitements améliorent la forme du résumé final.*

Le principe de la méthode est de fusionner les résumés soumis par plusieurs systèmes de résumé automatique ayant des caractéristiques différentes. Chaque système effectue un résumé par extraction à partir d'une segmentation en phrases commune à tous les systèmes. Seuls les identifiants de phrases sont véhiculés pour réunir les sorties de tous les systèmes dans un espace de représentation commun. La fusion est réalisée dans cet espace et implique à la fois la cohérence des sorties de chaque système et un certain nombre d'heuristiques destinées à améliorer la qualité linguistique des résumés.

[2]Les actes de DUC 2006 sont disponibles sur http://www-nlpir.nist.gov/projects/duc/pubs.html, visité en janvier 2007.

Les systèmes

Deux types de systèmes sont utilisés pour la sélection de phrases : des systèmes conçus pour la création de résumés multi-documents par extraction et des systèmes issus de la tâche question-réponse. L'idée est d'essayer de profiter des avantages de la spécialisation de chaque type de système à la fois pour déterminer globalement l'information importante issue de plusieurs documents et pour répondre aux questions précises accompagnant les *topics* DUC. Les 5 systèmes mis en œuvre sont décrits dans (Favre et al., 2006) et représentent un travail de collaboration entre les équipes du laboratoire.

- Le système S_1 est fondé sur les travaux présenté en 5.3 (résumé par MMR-LSA). Les phrases sont projetées dans un espace LSA construit à partir d'une matrice de cooccurrences sur le corpus DUC 2006[3]. Dans MMR[4], la fonction $\phi(\cdot)$ est réduite à un retrait des phrases de moins de 10 mots pleins. La fonction $\psi(\cdot)$ correspond $cosine(\cdot)$ dans l'espace LSA. La requête est interpolée[5] avec le centroïde des phrases pour en améliorer la portée. Le contexte des phrases est introduit par une interpolation de chacune d'entre elles avec la précédente[6] dans l'espace LSA ;
- le système S_2 a été construit à partir d'une version modifiée de CORTEX (Torres-Moreno et al., 2005) fondé sur différentes caractéristiques des phrases pour en guider la sélection (position, pertinence, redondance, longueur...) et une procédure de décision élaborée ;
- le système S_3 utilise une notion d'alignement des phrases avec les concepts exprimés dans le *topic*. Le score final d'une phrase est fonction de cet alignement à plusieurs niveaux morphologiques (stem, lemme, mot...), d'un score de couverture et de la position de la phrase dans le document ;
- le système S_4 est le module de recherche de passages pertinents du système de questions-réponses du LIA (Bellot et al., 2003). Le score d'une phrase est influencé par la densité d'apparition des mots du *topic* dans le document ;
- le système S_5, décrit dans (Gillard et al., 2005), est le module d'extraction de réponse du système de questions-réponses du LIA. Le score d'une phrase est exprimé à travers la « compacité » d'apparition des mots du *topic* dans un contexte proche autour de celle-ci.

Les systèmes S_1 et S_2 sont issus de la problématique « résumé automatique » et intègrent une minimisation de la redondance, alors que S_3, S_4 et S_5 répondent aux problématiques questions-réponses et se concentrent sur la précision de la réponse au *topic*. Il est intéressant de noter que S_4 et S_5 n'ont pas été modifiés pour la tâche DUC. Notamment, leurs paramètres n'ont pas été optimisés grâce aux données d'apprentissage. Cette dernière caractéristique permettra de déterminer si le corpus d'apprentissage apporte de meilleurs résultats sur le corpus de test malgré sa petite taille.

[3]La matrice est construite sur une fenêtre glissante de 30 mots pleins et représente les 60000 mots les plus fréquents par rapport aux 3000 mots les plus fréquents. La réduction de la matrice par SVD est fixée empiriquement à 200 dimensions.

[4]Les paramètres sont $\lambda_1 = 1$, $\lambda_2 = 0.95$, $\lambda_3 = 0.05$.

[5]Facteur d'interpolation avec le centroïde de 0.9.

[6]Facteur d'interpolation avec la phrase précédente de 0.05.

Traitements linguistiques

WASHINGTON (AP) –(1) **The**(2) Department of Housing and Urban Development is taking steps to preserve thousands of federally subsidized housing units for the poor. Housing Secretary Andrew Cuomo said Thursday the department will begin increasing payments to landlords who participate in the program. Many of the payments, set years ago, are below current market levels, and increasing them may entice landlords to stay in the program**, he said**(3). The increased payments will cost $30 million this fiscal year, HUD said...

<s> the Department of Housing and Urban Development is taking steps to preserve thousands of federally subsidized housing units for the poor . **</s>**(4)
<s> housing Secretary Andrew Cuomo said Thursday the department will begin increasing payments to landlords who participate in the program . </s>
<s> many of the payments **,**(5) set years ago , are below current market levels , and increasing them may entice landlords to stay in the program </s>
<s> the increased payments will cost $30 million this fiscal year </s>
...

FIG. 6.2: *Illustration des pré-traitements appliqués à un document DUC (APW19990428.0245). La marque de l'agence de presse est supprimée (1) ; les majuscules de début de phrase sont transformées en minuscules (2) ; les marques de discours rapporté sont supprimées (3) ; le document est segmenté en phrases (4) ; la ponctuation est séparée des mots (5).*

Les pré-traitements et post-traitements linguistiques améliorent le fond et la forme des résumés produits. Avant la phase de « sélection de phrases », un pré-traitement du texte brut aboutit au découpage en phrases et en mots communs utilisé par les systèmes de résumé (quelques pré-traitements sont illustrés par la figure 6.2). Cette étape est nécessaire afin de normaliser la morphologie des mots et de supprimer les éléments qui pourraient parasiter la modélisation de l'information contenue dans les phrases. Les traitements suivants sont réalisés :

- suppression des marque d'agence de presse (lieu, date et source) ;
- normalisation du vocabulaire selon un lexique propre ;
- découpage en phrases ;
- suppression des majuscules en début de phrase ;
- suppression des titres de personnes (Mr, Mme, Dr...) ;
- formatage des dates et quantités numériques ;
- suppression des formules rhétoriques organisant le discours ;
- nettoyage de la ponctuation (doublons, fins de phrase ...) ;
- suppression des expressions rapportant le discours d'une tierce entité ;
- suppression d'expressions temporelles relatives.

Tous ces traitements sont effectués grâce à un ensemble de règles, de dictionnaires et d'expressions régulières. L'objectif est d'avoir des phrases les plus proches possibles des phrases utilisées dans les résumés. Pour cela, les règles sont écrites de façon à réduire la longueur des phrases et minimiser le « risque linguistique » des pré-traitements. Hors contexte, les formules de construction du discours peuvent dégrader rapidement la cohérence du résumé, par exemple en opposant deux phrases qui ne traitent pas du même sujet. La suppression des expressions liées au discours rapporté (... a dit .., ...

écrit que ...) peut être polémique étant donné que la source de l'information, et donc sa crédibilité, est perdue. Ce type de règle permet toutefois de centrer les phrases sur leur thème principal tout en réduisant leur taille. Finalement, seules les règles qui amènent un résultat correct dans la majorité des cas sont conservées.

republics(1) of the former Soviet Union agreed in talks at Nato headquarters in Brussels to enforce reductions in heavy army weapons and aircraft as possible and without renegotiating the 1990 Conventional Arms Forces in Europe treaty(2) bush , responding to a series of Soviet proposals for reducing conventional forces , agreed for the first time to include manpower , helicopters and land based military aircraft in the Conventional Forces in Europe talks in Vienna . **bush , responding to a series of Soviet proposals for reducing conventional forces , agreed for the first time to include manpower , helicopters and land based military aircraft in the Conventional Forces in Europe talks in Vienna .**(3) while strategic nuclear arms will be the main topic of the trip , Baker also is expected during his four day stay in Moscow to provide greater detail about Bush 's new proposal to slash U.S. and Soviet troop strength in Central and Eastern Europe . the 22 nations represented at the Conventional Forces in Europe talks began cleaning up the 200 page text , which sets ceilings on the number of weapons each alliance can hold . **republics of the former Soviet Union agreed in talks at Nato headquarters in Brussels yesterday to enforce reductions in heavy army weapons and aircraft as soon as possible and without renegotiating the 1990 Conventional Arms Forces in Europe treaty , David White writes .**(3)

Republics of the former Soviet Union agreed in talks at Nato headquarters in Brussels to enforce reductions in heavy army weapons and aircraft as possible and without renegotiating the 1990 Conventional Arms Forces in Europe treaty. Bush, responding to a series of Soviet proposals for reducing conventional forces, agreed for the first time to include manpower, helicopters and land based military aircraft in the **Conventional Forces in Europe (CFE)**(4) talks in Vienna. While strategic nuclear arms will be the main topic of the trip, Baker is expected during his four day stay in Moscow to provide greater detail about Bush's new proposal to slash US and Soviet troop strength in Central and Eastern Europe. The 22 nations represented at the **CFE**(5) talks began cleaning up the 200 page text, which sets ceilings on the number of weapons each alliance can hold. **The successor states to the Soviet Union have promised to agree by the end of May on a share out of the weapons cuts to which Moscow committed itself under the 1990 CFE treaty. Gorbachev's announcement that he will propose massive cuts in military manpower on both sides came as part of a new, more detailed Soviet proposal for the current Vienna talks on CFE, US officials said.**(6)

FIG. 6.3: *Illustration des post-traitements appliqués à un résumé DUC (D398E). Les débuts de phrase sont recapitalisés (1) ; la ponctuation est normalisée et les phrases sont terminées par des points (2) ; les phrases dupliquées sont supprimées (3) ; les acronymes sont d'abord présentés avec leur forme complète (4), puis remplacés par leur forme réduite (5) ; la réduction du nombre de mots permet d'introduire de nouvelles phrases (6).*

Après formulation d'un résumé, des post-traitements sont appliqués dépendant de l'ordre des phrases ou pouvant malmener les systèmes de sélection de phrases (la figure 6.3 présente un résumé avant et après les post-traitements).

- Réécriture des acronymes ;
- réécriture des noms de personnes ;
- suppression des expressions entre parenthèses ;
- normalisation de la ponctuation pour maximiser le nombre de mots au sens de DUC (éléments séparés par des espaces).

Le principe de réécriture des acronymes et noms de personnes est le suivant : la première occurrence est complète et les suivantes utilisent une forme réduite. Pour les

acronymes, la forme complète détaille la signification de l'acronyme et la forme réduite est l'acronyme lui même. Pour les noms de personne, la forme complète contient le prénom et le nom alors que la forme réduite est limitée au nom de famille seul. Il n'est pas évident de détecter les noms et les acronymes et encore moins d'en faire la résolution lorsque la forme complète n'est pas connue.

Les définitions d'acronymes sont découvertes dans le corpus sous la forme d'une séquence de mots suivie d'un mot entre parenthèses. Les lettres de l'acronyme sont ensuite alignées sur la forme développée à l'aide de quelques heuristiques sur les majuscules, les déterminants et les conjonctions. Le score d'alignement, la fréquence d'occurrence et le nombre de résultats d'une requête jointe sur le moteur de recherche Google permettent d'établir un score de confiance pour ne garder que les résolutions les plus probables. Une expansion en aveugle sur Google ou un corpus plus volumineux est possible, mais comme les acronymes ont souvent plusieurs significations, la méthode permettant de choisir la bonne forme développée doit prendre en compte son contexte d'utilisation afin d'éviter les erreurs grossières comme celle présentée dans la table 6.2.

The test can be just as valuable if it discourages athletes from using European Patent Office (EPO) for fear they might ...

TAB. 6.2: *Exemple de résolution erronée de l'acronyme EPO à l'aide de Google. La bonne forme complète est Erythropoietin (EPO).*

La détection de noms de personnes est compliquée car il faut pouvoir différencier des noms de personnes utilisés pour représenter une marque (ou un lieu) de ceux représentants des personnes physiques. La présence de titres, de noms de métier, de prénoms et l'utilisation des formes étendues et réduites dans le corpus permettent d'établir une mesure de confiance dans la construction de la résolution. Toute la difficulté est de détecter par exemple la marque de cigarettes « Philip Morris » bien qu'elle contienne un prénom pour ne pas la remplacer par « Morris ».

Corpus DUC	Sans	Avec
2005	260.50 mots	249.26 mots
2006	259.00 mots	249.22 mots

TAB. 6.3: *Longueur moyenne des résumés avec et sans les traitements linguistiques. Ils permettent de réduire le nombre de mots d'environ 5%, ce qui correspond dans le cadre de la tâche DUC à ajouter une phrase au résumé.*

L'ensemble des traitements linguistiques améliorent la lisibilité de façon significative, bien que ce point reste difficile à mesurer. Ces traitement sont néanmoins très dépendants du type de données. La longueur des résumés est réduite d'environ 5%, soit 10 mots sur les 250 accordés dans la tâche DUC (table 6.3). Ceci représente une phrase supplémentaire en moyenne, apportant potentiellement un gain non négligeable de contenu dans un résumé.

Fusion

Les sorties des différents systèmes de sélection de phrases sont représentées dans un même espace de recherche afin de générer un résumé optimal. Le processus de fusion représente les contraintes du problème sous la forme d'un automate à états finis pondéré (*Weighted Finite State Transducer*, WFST) et le meilleur chemin dans cet automate est déterminé par programmation dynamique (Mohri et al., 2002). La construction de l'automate est la suivante :

1. chaque phrase est représentée par un transducteur acceptant ses mots en entrée et générant l'identifiant de la phrase en sortie ;

2. une seule occurrence de chaque phrase est conservée ;

3. les transducteurs sont concaténés en un unique transducteur et chaque phrase est doublée d'une transition « epsilon » pour autoriser n'importe quelle sélection de phrase ;

4. un automate est construit pour représenter un résumé valide d'environ 250 mots. Il contient 251 états et 250 transitions acceptant l'ensemble du vocabulaire, seuls les 20 derniers états sont finaux. Une fois composé avec le transducteur représentant les sélections de phrases possibles, seules les sélections aboutissant à une longueur entre 230 et 250 mots sont conservées.

5. le graphe d'hypothèses est finalement pondéré en utilisant une fonction de coût à plusieurs niveaux. Chaque phrase a un coût pondéré selon le nombre de systèmes pour lesquels la phrase est avant un rang donné et selon son rang maximum dans les sorties des systèmes. Un coût est associé aux états finaux pour avantager les résumés plus longs. Enfin, une pénalité est définie sur des mots spécifiques pour minimiser les anaphores pronominales en début de phrase et les références temporelles relatives.

La recherche du résumé optimal (chemin de coût minimal dans le graphe de fusion) est illustrée par l'équation 6.1. Dans cette équation, r est un résumé, p une phrase et m un mot ; $c_R(\cdot)$ est le coût associé à un résumé ; $c_P(\cdot)$ est le coût associé à une phrase ; $c_M(\cdot)$ est le coût associé à un mot ; l_{max} est la limite en nombre de mots d'un résumé ; $l(\cdot)$ représente sa longueur effective ; $rg_{sys}(p)$ est le rang de la phrase p pour le système sys ; $nb_{sys}\{rg_{sys}(p) < N\}$ représente le nombre de « votes » pour la phrase p. Les α_i sont

des hyperparamètres qui doivent être affinés sur un ensemble d'apprentissage[7].

$$\hat{r} = \underset{r}{\operatorname{argmin}}\, c_R(r) \tag{6.1}$$

$$c_R(r) = \alpha_1 \times (l_{max} - l(r)) + \sum_{p_i \in r} c_P(p_i)$$

$$c_P(p) = \alpha_2 \times \underset{sys}{\operatorname{nb}}\{rg_{sys}(p) < N\} + \alpha_3 \times \max_{sys} rg_{sys}(p) + \sum_{m_j \in p} c_M(m_j)$$

$$c_M(m) = \begin{cases} \alpha_4 & \text{si } m \text{ est une expression anaphorique en début de phrase} \\ \alpha_5 & \text{si } m \text{ est une expression anaphorique} \\ 0 & \text{sinon} \end{cases}$$

Les post-traitements appliqués après la fusion ont la particularité de modifier la longueur des phrases, avec pour conséquence que le résumé n'est plus optimal (en moyenne, dans le cas de DUC, une phrase peut y être ajoutée). Il faut réaliser une deuxième passe de fusion afin de prendre en compte les nouvelles longueurs de phrases, en forçant la présence des phrases modifiées sur lesquelles dépendent d'autres phrases du résumé (résolution d'acronymes, de noms de personnes).

Ordre chronologique et géographique

La campagne DUC 2005 a montré que l'organisation et la cohérence des résumés devaient être grandement améliorées. Pour cela, les phrases constituant chaque résumé de notre système sont ordonnées en fonction d'un ordre partiel temporel et géographique dépendant des caractéristiques du *topic* traité. Cette méthode repose sur l'observation que la plupart des *topics* DUC traitent d'événements ayant un étalement temporel important (plusieurs années) et impliquant plusieurs régions au niveau mondial.

Afin de déterminer le type d'ordre appliqué à un *topic*, des règles simples permettent de lui donner les étiquettes suivantes : spécifique, général, temporel, et géographique. Par exemple, un *topic* est étiqueté comme général s'il ne contient pas de noms propres ; un *topic* est étiqueté comme géographique s'il contient des mots comme « pays », « mondial », ou « nation » ; un *topic* est étiqueté comme temporel s'il contient des mots comme « événements », ou « dates ».

Chaque document se voit attribué une étiquette temporelle (l'année de publication de la nouvelle) et une étiquette géographique (le pays de publication). Les phrases sont d'abord ordonnées selon les étiquettes de leur document d'origine, puis selon leur position à l'intérieur du document. Les deux sous-ordres sont appliqués séquentiellement en fonction des caractéristiques du *topic* (les règles sont déterminées empiriquement sur le corpus de développement). Le premier ordre permet de former des paragraphes éventuellement introduits par l'étiquette en question. Des exemples sont donnés dans la table 6.4.

[7] Pour DUC 2006, les paramètres sont $N = 30$, $\alpha_1 = 1$, $\alpha_2 = 1000$, $\alpha_3 = 100$, $\alpha_4 = 2$, $\alpha_5 = 1$. Les expressions anaphoriques pénalisantes sont les pronoms et les jours de la semaine.

Topic : NASA's Galileo Mission
How successful was NASA's Galileo space probe mission of Jupiter ? What discoveries were made about the **planet*** and its moons ? Include details about **when*** the probe was launched and any troubles it may have encountered.

Résumé généré :
In London, in 1996, the planet Jupiter is much hotter and windier than previously believed, the latest information from the space probe Galileo has shown.
In Washington, in 1998, volcanoes on a moon of Jupiter called Io may be the hottest place in the solar system outside the sun itself. Using instruments orbiting Jupiter on the Galileo spacecraft, researchers calculated the temperature of lava spewing from volcanoes on Io. In addition to the gravitational studies, Galileo has measured magnetic fields surrounding Jupiter and its moons. For about two years the space probe Galileo has been gathering ever more evidence that a large ocean lies hidden beneath the frozen, fractured surface of Jupiter's moon Europa. Photographs taken by the Galileo spacecraft show that Jupiter's moon Io is aglow with colorful light.
In Pasadena, in 1999, the Galileo spacecraft halted all non essential activities by going into a safe mode shortly after close approaches to the moon Europa and Jupiter.
In Los Angeles, NASA's aging Galileo spacecraft flew within 380 miles of Jupiter's moon Io, exposing the craft to so much radiation that mission controllers feared the probe might not survive. The aging and glitch prone Galileo spacecraft successfully flew within 380 miles of Jupiter's moon Io, overcoming huge doses of radiation and a computer problem just hours before the approach.
In Pasadena, in 2000, NASA's aging Galileo spacecraft swooped past Jupiter's frozen moon Europa on Monday but apparently did not experience any computer problems from the planet's intense radiation.

FIG. *6.4:* *Exemple de topic (sujet) classé par erreur à la fois comme temporel et géographique (D0638B). Les indices ayant provoqué cette classification sont dénotés par une étoile*. Les paragraphes du résumé sont introduits par le lieu géographique du document d'origine et son année de publication. Malgré leur bon potentiel pour améliorer la structure des résumés, ces indications représentent un risque d'incohérence avec les informations déjà présentes dans les phrases et avec la structure de la phrase (mauvais temps des verbes...).*

Les étiquettes géographiques et temporelles sont détectées au niveau du document et non de la phrase. Cette méthode peut faire émerger des informations contradictoires avec le contenu de la phrase. Pour limiter ce risque, les paragraphes contenant déjà des références temporelles ou géographiques ne sont pas introduits par l'étiquette correspondant au document.

6.1.2 Résultats sur DUC 2006

L'édition 2006 a impliqué 34 participants évalués entre eux et par rapport au système trivial NIST générant des résumés à partir des 250 premiers mots du document le plus récent parmi les documents correspondant à un *topic* donné. Le protocole et les mesures d'évaluation ont été décrits dans la section 2.2.1. La figure 6.5 contient les résultats globaux du système LIA-Thales sur l'ensemble des évaluations manuelles et automatiques, accompagnées du classement du système par rapport aux autres systèmes. Notre soumission obtient un très bon classement sur les scores prenant en compte le fond des résumés, compte tenu du peu d'expérience de l'équipe sur DUC (première participation). Par contre, dans l'absolu, on observe que le Contenu et la Qualité générale représentent 50% du score idéal. Au niveau des évaluations automatiques, environ

En 1998, paragraphe ordonné géographiquement.
En 2001, paragraphe ordonné géographiquement.
En 2004, paragraphe ordonné géographiquement.
Aux États Unis, paragraphe ordonné temporellement.
Au Mexique, paragraphe ordonné temporellement.
En Amérique du sud, paragraphe ordonné temporellement.

TAB. 6.4: *Ordre des phrases selon les caractéristiques du topic. Dans le premier cas, l'ordre temporel prévaut alors que dans le second, l'ordre géographique prévaut. Dans les 2 cas, une étiquette introductive est générée.*

Évaluation manuelle	**Score**	**Rang /35**	**Min**	**Max**
Qualité linguistique moyenne	3.57	14	2.32	4.08
Grammaticalité	4.08	7	1.38	4.62
Non redondance	3.84	31	3.76	4.66
Clarté des références	3.42	6	1.90	4.00
Focalisation	3.74	13	2.50	4.28
Structure et cohérence	2.76	19	1.16	3.28
Contenu	2.80	8	1.68	3.08
Qualité générale	2.40	8	1.34	2.84
Pyramids	0.21	6/21	0.13	0.25
Évaluation automatique	**Score**	**Rang /35**	**Min**	**Max**
Rouge 2	0.08700	6	0.02834	0.09558
Rouge SU4	0.14522	5	0.06394	0.15529
Basic Elements	0.04763	6	0.00459	0.05107

TAB. 6.5: *Résultats du système LIA-Thales sur DUC 2006.*

8% des bigrammes de l'ensemble des résumés de référence apparaissent dans la soumission du système. La qualité linguistique du système est correcte avec un rang de 14 sur 35. Une analyse détaillée fait apparaître que la redondance de surface (utilisation de pronoms à la place de la forme complète) est relativement mauvaise comparée aux autres systèmes. Cela s'explique par la pénalisation des phrases contenant des pronoms dans le processus de fusion. Cependant, une conséquence de ce traitement est l'obtention d'un bon score de clarté des références. Il faut tout de même remarquer que les résumés sont relativement peu focalisés et que malgré les efforts pour les structurer, le score de structure et cohérence reste relativement bas.

Un exemple de résumé généré par le système LIA-Thales est présenté dans la figure 6.6 conjointement à un des résumés de référence. La table 6.8 donne les différents scores d'évaluation associés à ce résumé : ce *topic* a obtenu de bons scores dans la plupart des domaines et représente un bon exemple de conditions de bon fonctionnement du système.

La figure 6.5 montre les performances Rouge 2 et Rouge SU4 (pour DUC 2005 et DUC 2006) des différents sous-systèmes avant fusion (S_1 à S_5), la fusion des 3 meilleurs

Soumission	Scientists looking for signs of global warming should spend more time scrutinizing Earth's weather circulation patterns, a new study suggests. The findings suggest that global warming blamed on human industry is manifesting itself not just by causing a global rise in temperatures but by favoring certain established weather patterns. While the study does not prove that human induced global warming is what caused the change in frequency of circulation patterns, he said, the change is consistent with it. The theory is that at certain critical altitudes, roughly from 6 to 12 miles, increasing carbon dioxide has the effect of warming the tropics but cooling the polar regions. The data suggest that global warming has caused a slowdown of the Earth at a rate of 0.56 milliseconds a century.. This warming would touch off widespread disruptions in climate and weather and cause the global sea level to rise and flood many places. In Beijing, chinese scientists believe that global warming, particularly higher winter temperatures, will continue to the first half of the 21 st century. In San Antonio, while the Global Climate Coalition questions whether global warming is happening, it advocates that companies voluntarily explore and employ new technology to reduce emissions that contribute to global warming. In Washington, greenhouse gas emissions blamed for global warming may cause the collapse of the West Antarctic Ice Sheet and raise the average global sea level by four to six metres, beginning as as the, a new scientific study predicted recently.
Référence	Global warming is thought to be at least partly caused by emissions of waste industrial gases like carbon dioxide, produced by burning fossil fuels like coal, oil and natural gas. These emissions trap solar radiation and produce a greenhouse effect. Methane and nitrous oxide emissions from agriculture (ruminants and manure) make up 8% of greenhouse gases. Controls on sulfur dioxide emissions reduce a balancing cooling effect. Global warming already causes more frequent El Nino appearances, receding shorelines, longer warm seasons, and a slower earth spin. It affects habitats and threatens marine life. If emissions are not reduced, average surface temperature will rise 2-6 degrees over the next century, bringing widespread climatic, ecological and economic dislocation. Floods and droughts will increase in frequency and intensity. Melting polar ice will cause rising sea levels and coastal flooding. Malaria will increase. Rates of habitat loss and species extinction will increase. Communities will need to adapt to new conditions. Skeptics argue that human activities have little influence on climate. Most observed warming is due to natural causes like changes in solar radiation or the circulation of heat-bearing ocean waters. Measurements taken by satellites have found little temperature rise in the upper atmosphere. Computer models are unreliable. Any warming over the next century would be most pronounced in the winter, at night, and in sub-Arctic regions, doing little harm and creating benefits like longer growing seasons and faster plant growth. Industry argues that reducing the use of fossil fuels would cause economic harm to consumers.

TAB. 6.6: *Soumission du système et l'un des résumés de référence pour le topic D0641 (Global warming) : certains paragraphes ont été introduits par le lieu géographique concerné. Les quelques erreurs de post-traitements montrent les limites d'une approche à base de règles. Bien que ce résumé soit l'un des meilleurs produits par le système, la qualité de l'abstraction n'est pas encore à la hauteur de ce que produit un expert humain. Un comparatif des résumés produits par les 5 sous-systèmes sur ce topic est disponible en annexe A.*

systèmes ($F1$) et de l'ensemble des systèmes ($F2$). Les mesures d'évaluation permettent de comparer le comportement des sous-systèmes sur les données de développement (2005) et de test (2006). Les paramètres des systèmes S_1, S_2 et S_3 ont été optimisés dans le but de maximiser les scores Rouge sur 2005, contrairement à S_4 et S_5 qui ont été

Soumission

The custody flap over 6 year old Elian Gonzales could ultimately strengthen U.S. Cuba relations if American officials stand firm and do not succumb to political pressure. The organizers cited a more positive course in the ongoing custody battle over the 6 year old. Dan Burton, R Ind., to have the boy testify before a House committee and efforts by Elian's relatives to attain custody of the child. As the city's streets settled into an uneasy calm, the battle over six year old Cuban rafter Elian Gonzalez moved to court Friday with the boy's Miami relatives hoping to overturn a decision by federal immigration officials to send him back to his father in Cuba. As Gonzalez was speaking, his relatives in Miami appeared in a county Family Court, with Elian's great uncle, Lazaro Gonzales, asking for temporary custody of the boy. Cuba ELIAN miami demonstrators sang and prayed outside the home where Elian Gonzales is staying as all sides in the custody battle waited for a federal appeals court ruling that could lead to the boy's reunion with his father. Cuba ELIAN miami as both sides in the Elian Gonzales custody case await a federal appeals court ruling, the mayor of Miami flies to Washington to meet with Attorney General Janet Reno. Cuba ELIAN miami pressure mounts on the relatives of Elian Gonzales to turn the boy over to his father as Attorney General Janet Reno reportedly has given approval to take the child by force, if necessary.

TAB. 6.7: *Exemple de résumé généré par notre système pour le topic D0647 (Polémique autour de l'enlèvement d'Elian Gonzales). Le résumé a une très faible qualité linguistique a cause d'une mauvaise détection des indications de lieu et source en début de document. Toutefois, les scores automatiques de ce résumé sont élevés, en contradiction avec l'évaluation manuelle, ce qui montre les limites de l'évaluation Rouge.*

utilisés tels quels. Cela se traduit par une différence de performances sur ce jeu de données. Le fait que cette différence soit toujours significative sur le corpus 2006 indique que l'utilisation de données d'apprentissage est bénéfique, même si elles sont peu nombreuses (50 *topics*). Un autre aspect intéressant est que les différentes fusions ($F1$ et $F2$) ont de meilleurs résultats que le meilleur système, que ce soit pour 2005 ou 2006. Enfin, la fusion $F2$ (celle qui a été utilisée pour la soumission) admet de meilleurs résultats sur 2006 par rapport à $F1$ alors que ses résultats étaient moins bons sur 2005. Ceci prouve que l'utilisation d'un plus grand nombre de systèmes aux performances variées agit comme un garde-fou en limitant le sur-apprentissage presque inévitable sur des corpus de petite taille comme ceux de DUC.

Les résultats peuvent être observés en fonction des étiquettes données à chaque *topic*, qui ont un impact sur le choix de l'ordre des phrases et la façon d'introduire les paragraphes. Quatre étiquettes sont retenues selon que le *topic* est *Spécifique*, *Général*, *Géographique* ou *Temporel*. L'étiquette fictive *Inconnu* est ajoutée lorsque le *topic* n'est ni *Géographique*, ni *Temporel*. Comme le montre la table 6.10, les *topics* du type *Géographique* et *Général* obtiennent les meilleures performances. Visiblement, le fait d'introduire les paragraphes par le lieu géographique concerné améliore la contextualisation des phrases dans les résumés. Par contre, l'ajout d'un contexte temporel aux *topics* étiquetés *Temporel* n'apporte rien comparé aux *topics* étiquetés *Inconnu*.

Une étude des intervalles de significativité des évaluations automatiques dénombre les systèmes significativement meilleurs (à 95%) ou moins bon qu'un système donné. La figure 6.9 montre ces résultats pour le système LIA-Thales. Cette étude montre que le système est au niveau des meilleurs systèmes sur Rouge 2 et Rouge SU4.

Évaluation manuelle	**D0641**	**D0647**
Qualité linguistique moyenne	4.4	1.6
Grammaticalité	4	1
Non-redondance	5	2
Clarté des références	4	1
Focalisation	5	3
Structure et cohérence	4	1
Contenu	3	1
Qualité générale	3	1
Pyramids	n/a	0.083
Évaluation automatique	**D0641**	**D0647**
Rouge 2	0.08564	0.11011
Rouge SU4	0.14662	0.16382
Basic Elements	0.05161	0.05284

TAB. 6.8: *Résultats de l'évaluation pour les topics D0641 et D0647. Le premier est un des meilleurs résumés de la soumission alors que le second est le moins bon. Il faut remarquer que les scores automatiques du topic D0647 ne sont pas en accord avec les scores manuels.*

Évaluation automatique	**score**	**inf.**	**sup.**	**nb.** $>$	**nb.** $<$
Rouge 2	0.08700	0.00368	0.00396	2	25
Rouge SU4	0.14522	0.00365	0.00358	1	26
Basic Elements	0.04763	0.00299	0.00282	2	26

TAB. 6.9: *Évaluations automatiques du système LIA-Thales pour DUC 2006, avec les incertitudes inférieure (inf.) et supérieure (sup.) de chaque score et le nombre de systèmes significativement meilleurs (nb. $>$) et moins bons (nb. $<$) sur chaque score.*

Les résultats du système pour cette campagne d'évaluation montrent que le système S_1 (développé dans ces travaux) est au niveau de l'état de l'art et peut éventuellement être utilisé conjointement à d'autres systèmes grâce à un processus de fusion (figure 6.6). Malheureusement, les données d'évaluation sont textuelles et ne reflètent que partiellement les problématiques du résumé audio. Pour cela, la prochaine section est dédiée à la dégradation des données DUC pour simuler une structuration automatique d'un contenu parlé.

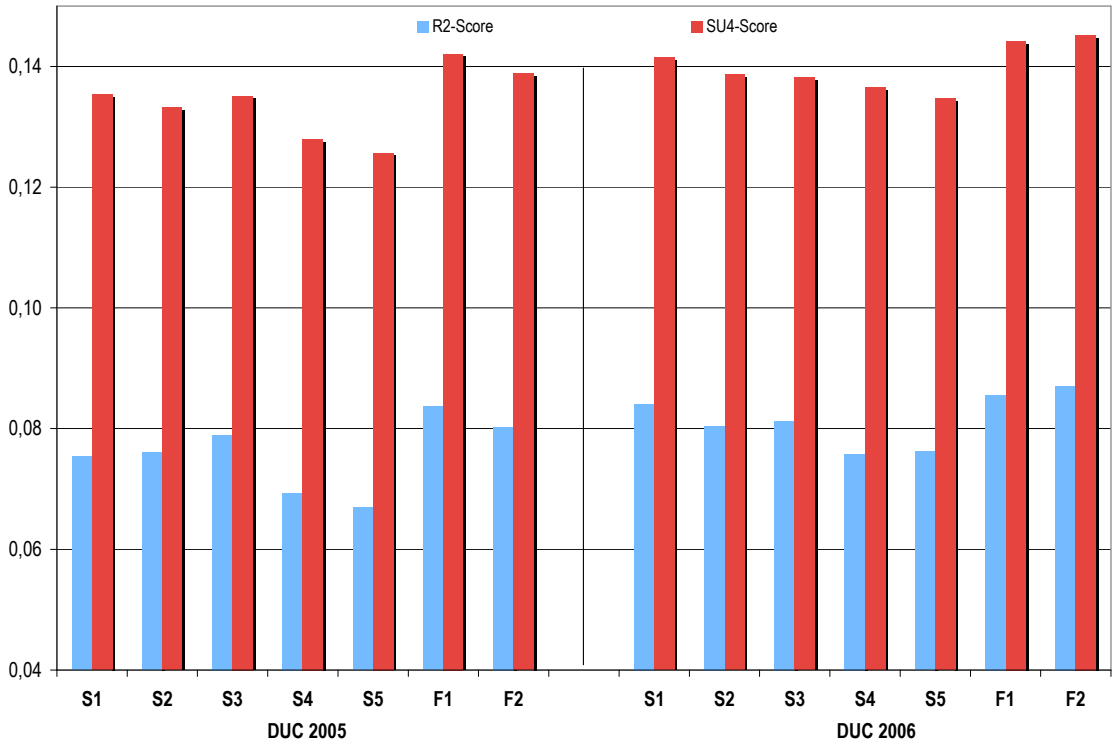

FIG. 6.5: *Scores Rouge 2 et Rouge SU4 sur les données de DUC 2005 et DUC 2006 pour les 5 systèmes LIA-Thales (S_1 à S_5), la fusion des 3 meilleurs (F1) et la fusion des 5 systèmes (F2). L'apprentissage sur le corpus 2005 est bénéfique car les systèmes non optimisés restent significativement moins performants que les autres sur DUC 2006. La fusion limite le sur-apprentissage lorsqu'elle est appliquée sur les 5 systèmes (F2 est meilleure sur DUC 2006).*

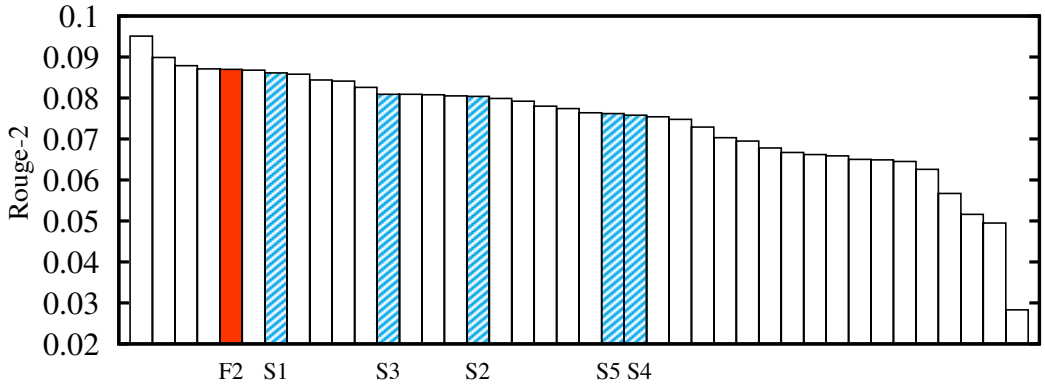

FIG. 6.6: *Classement des sous-systèmes S_1 à S_5 (barres hachées bleues) et de leur fusion F_2 (barre pleine rouge) par rapport aux soumissions des autres participants à DUC 2006. Les scores sont exprimés selon la mesure Rouge-2. Le système S_1 est à la hauteur des systèmes état de l'art en résumé automatique de texte.*

Évaluation manuelle	**Spécif.**	**Gén.**	**Temp.**	**Géo.**	**Inconnu**
Qualité linguistique moyenne	3.39	**3.58**	3.34	**3.77**	3.48
Grammaticalité	3.90	**4.30**	3.80	4.00	**4.25**
Non redondance	3.52	**4.28**	3.81	**4.33**	3.75
Clarté des références	**3.59**	3.19	**3.69**	3.33	3.28
Focalisation	**3.79**	3.66	3.37	**4.33**	3.82
Structure et cohérence	2.14	**2.42**	2.00	**2.83**	2.28
Contenu	2.55	**3.09**	**3.06**	3.00	2.57
Qualité générale	2.38	**2.48**	2.37	**2.66**	2.39
Évaluation automatique	**Spécif.**	**Gén.**	**Temp.**	**Géo.**	**Inconnu**
Rouge 2	**0.090**	0.082	0.077	**0.095**	0.091
Rouge SU4	**0.146**	0.144	0.136	**0.151**	0.149
Basic Elements	0.045	**0.051**	0.046	**0.056**	0.047

TAB. *6.10:* *Scores automatiques et manuels sur DUC 2006 en fonction des étiquettes, générées pour chaque topic, qui influencent l'ordre des phrases du résumé. Les étiquettes sont Spécifique (Spécif.), Général (Gén.), Temporel (Temp.), Géographique (Géo.) et Inconnu.*

6.2 Simulation de l'impact d'un contenu parlé

Les expériences présentées dans cette section visent à simuler l'impact des différents éléments de structuration automatique d'un contenu parlé sur la tâche DUC. Nous nous attachons à évaluer le fond du résumé par la méthode automatique Rouge-2 tout en ignorant la forme car aucune méthode automatique ne permet de le faire. La chaîne de structuration du contenu audio décrite en 3.1 peut amener divers types d'erreurs :

1. la détection des classes acoustiques peut pousser à ignorer des segments entiers de parole par confusion avec la classe musique, par exemple. Ces segments ne pourront pas être utilisés dans le résumé ;
2. la segmentation en locuteurs a un impact sur la sélection des modèles de transcription et sur la segmentation en phrases. Les erreurs sur les tours de parole réduisent la qualité des frontières de phrase et celles sur les identités ont tendance à faire augmenter le taux d'erreur de mots ;
3. la transcription automatique engendre des erreurs sur le contenu linguistique sous la forme d'insertions, de substitutions et de suppressions de mots. Ces erreurs ont un impact sur l'ensemble des tâches de plus haut niveau, dont le résumé ;
4. l'extraction des entités nommées (comme n'importe quel type de descripteur sémantique) peut affecter une méthode de création du résumé les utilisant fortement ;
5. la segmentation en phrases est cruciale pour le rendu final du résumé audio car la qualité des syntagmes grammaticaux extraits est directement liée à la compréhension du contenu.

Les données traitées pour cette expérience sont celles de DUC 2006. Le système utilisé est le système S_1 de la soumission LIA-Thales décrit en 6.1.1. Les post-traitements appliqués ignorent toutefois le processus de fusion et ne changent en aucun cas l'ordre des phrases.

6.2.1 Cadre expérimental

L'ensemble des types d'erreurs provoqués par la structuration sont simulés sous la forme d'erreurs touchant la transcription : suppressions, insertions et substitutions du contenu lexical. Les autres types d'erreurs peuvent être inférés à partir de ces erreurs (par exemple, la suppression d'un segment correspond à une suppression en chaîne des mots). Obtenir une dégradation réaliste pour émuler le comportement d'un système de transcription est envisageable (Deng et al., 2003), mais très dépendant de nombreux paramètres liés aux données, à l'implémentation et à la tâche. Les dégradations appliquées dans les expériences suivantes sont uniformes selon l'algorithme de la figure 6.7. Malgré ses défauts, une dégradation uniforme a l'avantage de représenter le cas le plus défavorable pour un système de sélection de phrase car ce dernier ne pourra pas profiter de la variance des dégradations pour choisir des phrases plus « propres ». Des exemples de phrases dégradées sont donnés dans la table 6.11.

pour chaque *mot* $\in$ *documents* **faire**
  p =nombre au hasard uniforme sur $\in [0;1]$;
  si $p < p_{sup}$ **alors**
    supprimer le mot;
  sinon si $p < p_{del} + p_{ins}$ **alors**
    insérer un mot du vocabulaire choisi au hasard;
  sinon si $p < p_{del} + p_{ins} + p_{sub}$ **alors**
    substituer le mot avec un mot choisi au hasard;
  sinon
    utiliser le mot d'origine;
  fin
fin

FIG. 6.7: *Algorithme pour la génération aléatoire des erreurs selon une probabilité de suppression (p_{del}), d'insertion (p_{ins}) ou de substitution (p_{sub}). Le taux d'erreur de mots obtenu n'est pas exactement de $p_{del} + p_{ins} + p_{sub}$ à cause des effets de bord entre les différents types d'erreurs.*

Pour mieux comprendre l'impact de données parlées, le système MMR-LSA est testé selon deux conditions correspondant à un résumé textuel « lu » par l'utilisateur et à un résumé audio « écouté ». Pour mettre en œuvre la première condition, les données sont dégradées avant d'être injectées dans le système de résumé, dont les sorties sont utilisées pour calculer Rouge-2. Dans la seconde condition, l'utilisateur ne perçoit pas les erreurs de transcription. Cette contrainte est modélisée en remplaçant les phrases des résumés générés pour la condition précédente par les phrases propres d'origine. Ainsi, il est possible de mesurer directement la robustesse de la méthode de sélection de phrases.

Type	**Exemple**
Original	Andrew Cuomo said the department will begin increasing payments
Insert.	Andrew **lake** Cuomo said the department **change** will begin increasing payments **London**
Supp.	Cuomo the department will begin increasing
Subst.	Andrew Cuomo **lake** the department **London** begin increasing **change**
Tout-type	Andrew Cuomo **lake** the will begin **London** increasing payments
supp. EN	said the department will begin increasing payments
remp. EN	**lake change** said the department will begin increasing payments

TAB. 6.11: *Exemples de phrases dégradées. Les phrases ayant subit des insertions (Inser.), des suppressions (Suppr.) des substitutions (Subst.) et un mélange de ces trois types d'erreur (Tout type), ont un WER de 33%. Les phrases dont les entités nommées sont supprimées (suppr. EN) ou remplacées aléatoirement (remp. EN), ont un WER de 22%.*

Type de dégradation	WER	Rouge-2 « lu »		Rouge-2 « écouté »	
Aucune	0.0	0.08407		0.08407	
Remplacement des OOV	1.0	0.08255	-1.8%	0.08318	-1.0%
└ *écart-type*		*0.00034*		*0.00034*	
Suppression des OOV	1.0	0.08283	-1.4%	0.08279	-1.5%
Remplacement des EN	10.4	0.06741	-19.8%	0.08029	-4.4%
└ *écart-type*		*0.00083*		*0.00094*	
Suppression des EN	10.4	0.07211	-14.2%	0.07991	-4.9%
Erreurs aléatoires	10.0	0.07440	-11.5%	0.08232	-2.0%
└ *écart-type*		*0.00118*		*0.00104*	

TAB. 6.12: *Impact d'une transcription automatique sur les performances de MMR-LSA en dégradant artificiellement les documents de DUC 2006. Rouge-2 est calculé sur les phrases dégradées (condition d'un résumé « lu ») et sur ces mêmes phrases remplacées dans les résumés par leur contenu d'origine (condition d'un résumé « écouté »). Le vocabulaire est limité aux 65000 mots les plus fréquents de Gigaword pour simuler l'utilisation d'un lexique de cette taille. Les mots hors vocabulaire (OOV) sont soit supprimés, soit remplacés par des mots choisis aléatoirement dans le lexique. Les mêmes types de dégradations sont appliqués aux entités nommées (EN). L'impact sur les entités nommées est comparé avec celui de l'introduction d'erreurs aléatoires au même niveau de WER. Les expériences aléatoires sont répétées 50 fois.*

6.2.2 Résultats sur les données dégradées

Les documents ont d'abord été dégradés en limitant leur vocabulaire aux mots contenus dans un lexique de système de transcription. Ce lexique est construit en conservant les 65000 mots les plus fréquents d'un grand corpus, Gigaword (Graff, 2003), habituellement utilisé pour créer des modèles de langage. Cela correspond sur les données de DUC 2006 à un taux de mots hors-vocabulaire (OOV) de 1%. Cette limitation du lexique donne lieu à deux dégradations distinctes : le remplacement aléatoire des mots hors-vocabulaire et leur suppression. Ce type de lexique n'est pas forcément représentatif de la différence observée entre les données d'apprentissage et les données de test (essentiellement pour les modèles de transcription, provoquant une forte baisse des performances en test). Dans le cadre de journaux radio-diffusés, les entités nommées sont les plus touchées par cette différence. Nous explorons cet aspect en dégradant uniquement les entités nommées (suppression et remplacement). Ce type de dégradation est comparé à des erreurs aléatoires caractérisées par un taux d'erreur de mots similaire. La table 6.12 donne une idée de l'impact de toutes ces dégradations sur Rouge-2 dans les deux conditions que nous avons envisagées (« lu », « écouté »). Une analyse de ces résultats montre que l'impact sur Rouge-2 est toujours plus fort sur la condition « lu » que sur la condition « écouté ». La limitation du lexique aux mots les plus fréquents implique une baisse relativement faible de Rouge-2 : ce facteur n'est pas limitant mais correspond à des conditions optimales de fonctionnement de la transcription. Par contre, de manière attendue, la suppression des entités nommées réduit fortement la qualité des résumés lus (-14% sur Rouge-2) et a un impact similaire sur les résumés écoutés (-5% sur Rouge-2). Comparé à des erreurs aléatoires, ce type de dégradation est beaucoup plus pénalisant pour le résumé et prouve une nouvelle fois l'intérêt des

entités nommées.

Les figures 6.8 et 6.9 illustrent la variation de Rouge-2 par rapport au taux d'erreur de mots introduit par les différents types de dégradation. Les paramétrages de l'algorithme (fig. 6.7) aboutissent aux dégradations suivantes : uniquement des suppressions, uniquement des insertions, uniquement des substitutions et une distribution uniforme de ces trois classes d'erreurs. Les performances du système sont comparées à un résumé trivial par sélection aléatoire des phrases[8] (indépendante du système) et à un reclassement aléatoire des phrases du système avant sélection[9] (dépendant du système).

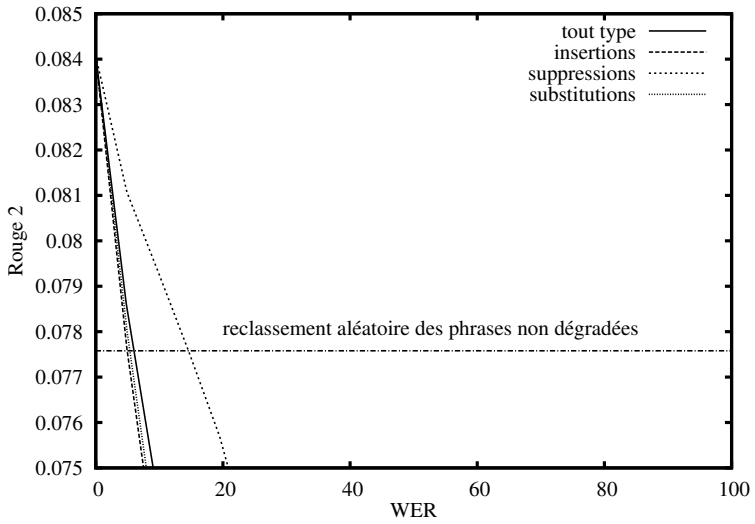

FIG. 6.8: *Comparaison des dégradations pour un résumé « lu », pour le système MMR-LSA sur les données DUC 2006. Les erreurs sont des suppressions, des insertions et des substitutions de mots dispensées de façon uniforme pour atteindre divers degrés de WER. Comme les erreurs se répercutent sur le texte du résumé, leur impact est très corrélé à la dégradation pour une mesure fondée sur la distribution des mots comme Rouge-2. Seules les suppressions sont compensées dans une moindre mesure par MMR-LSA. Cette figure est à la même échelle que la figure 6.9 présentant les résultats dans la condition « écouté » (ce choix d'échelle provoque le chevauchement des courbes).*

L'analyse de la figure 6.8 montre que Rouge-2 décroît fortement avec l'augmentation du WER lorsque le résumé est « lu ». Cette observation est prévisible compte tenu de la manière dont fonctionne Rouge. La proportion d'erreur des phrases se retrouve dans les résumés car les erreurs sont uniformes et chacun des *n*-grammes observés par Rouge a autant de chances d'être affecté que les autres. Les performances du reclassement aléatoire sont atteintes pour un WER d'environ 10% et celles de la sélection aléatoire

[8]Cette *baseline* est obtenue à partir des documents d'origine sans pré/post-traitements. Les phrases sont arbitrairement segmentées à chaque occurrence d'un point suivit d'un espace (« . »). Les performances obtenues (Rouge-2 de 0.05576) sont similaire à la *baseline* DUC consistant à créer un résumé à l'aide des 250 premiers mots du document le plus récent (Rouge-2 de 0.0495).

[9]Pré/post-traitements compris, sans appliquer d'autre dégradation (Rouge-2 de 0.07611).

pour un WER d'environ 30%. Seules les suppressions semblent être compensées par le système de résumé, quand l'algorithme choisit, par exemple, des phrases plus longues à l'origine que celles choisies sans dégradation. Ces diverses remarques montrent bien que, malgré les observations (dans des conditions réelles) de taux d'erreur de mots réduits dans les résumés par Christensen et al. (2003) et Murray et al. (2005), le résumé de parole sous forme textuelle nécessite de développer des techniques pour détecter et écarter les phrases mal transcrites.

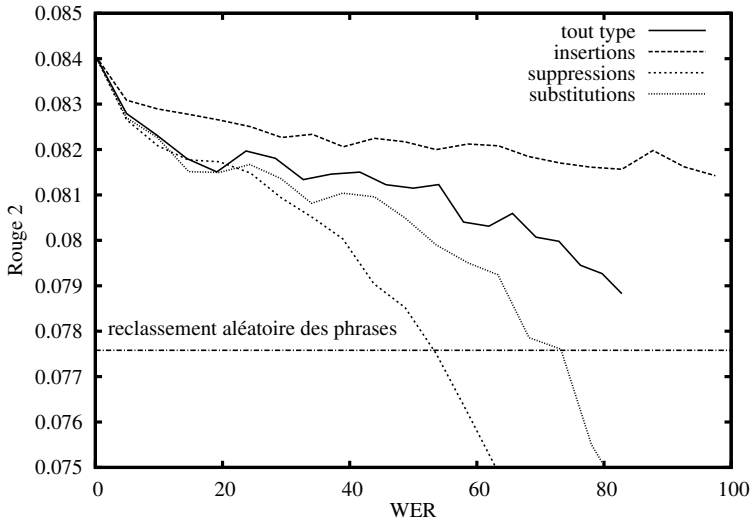

FIG. 6.9: *Comparaison des dégradations d'un résumé « écouté », pour le système MMR-LSA sur les données DUC 2006. Contrairement à l'expérience présentée par la figure 6.8, les phrases dégradées sont remplacées dans le résumé par leur version propre pour simuler une écoute de l'audio. Ceci amène à une bonne conservation des performances, même à de forts taux d'erreur, prouvant soit que MMR-LSA est robuste aux erreurs de ce type, soit que Rouge-2 ne reflète pas la qualité des résumés dans de telles conditions.*

La figure 6.9 est plus intéressante car MMR-LSA conserve des scores Rouge-2 élevés lorsque le résumé est « écouté », même sur des données fortement dégradées. Le système est globalement meilleur que les deux *baselines* aléatoires pour un WER inférieur à 50. De plus, les insertions ne sont jamais pénalisantes dans la mesure où le contenu d'origine est toujours présent dans les documents. Une analyse de la variance des résultats pour 50 initialisations différentes du générateur pseudo-aléatoire est présentée dans la figure 6.10. Bien que dans certaines conditions, les données aléatoires augmentent les performances du système (ce qui représente le défaut d'optimisation des paramètres du système sur DUC 2005), l'allure générale des courbes est représentative. Ces observations ne peuvent être expliquées que par une robustesse du système ou un échec de Rouge à évaluer des résumés « écoutés ».

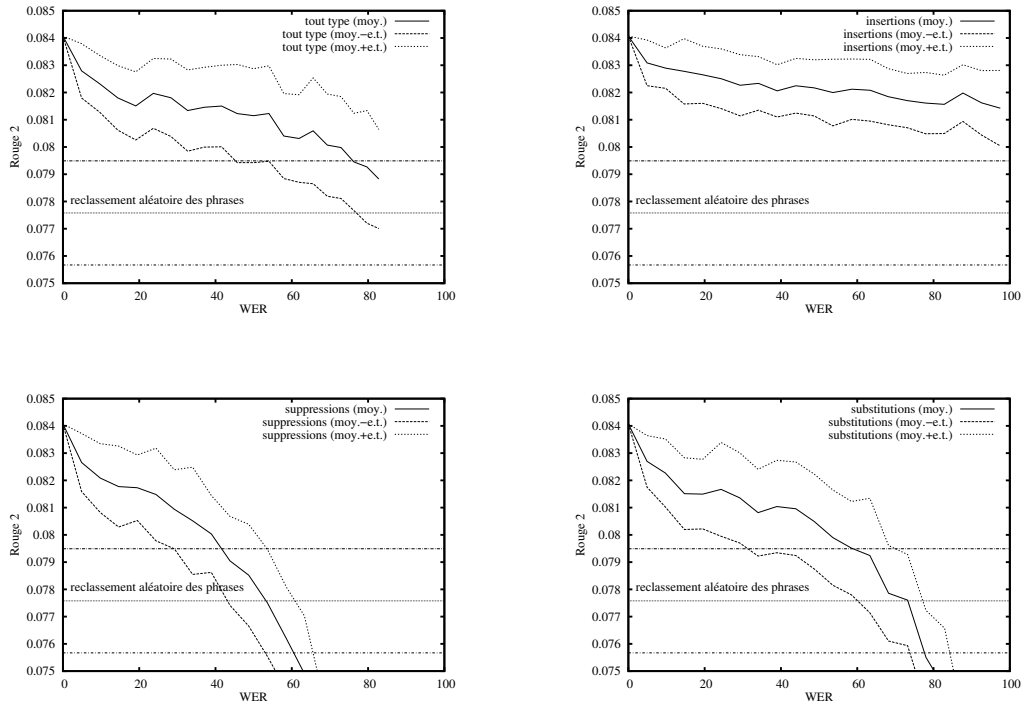

FIG. *6.10:* *Détail des courbes de performances Rouge-2 du système sur des données bruitées, dans la condition « résumé écouté ». Les performances sont comparées au reclassement aléatoire des sorties du système représentant une dégradation maximum de la sélection de phrases (Rouge-2 : 0.07611 de moyenne et un écart-type de 0.00191). La moyenne et l'écart-type sont illustrés pour chaque type d'erreurs. Les expériences aléatoires sont répétées 50 fois avec des initialisation différentes du générateur pseudo-aléatoire.*

6.2.3 Interprétation des résultats

La robustesse du système est obtenue principalement par l'utilisation du document moyen dans l'expression du besoin utilisateur (les mots les plus fréquents restent fréquents après dégradation). Ainsi, le bruit introduit dans les phrases est également présent dans la requête. Par exemple, pour une dégradation par insertion, le contenu originel des phrases est toujours présent et porteur d'une information plus cohérente que le bruit rajouté. Le reclassement aléatoire des sorties du système montre que les prétraitements et les post-traitements jouent aussi un rôle pour compenser le bruit. La composante la plus déterminante du pré-traitement consiste à écarter les phrases de moins de 10 mots informatifs. La longueur des phrases est prise en compte implicitement de cette façon et il a été prouvé que ce paramètre est déterminant pour le résumé. Le post-traitement contient un garde-fou pour éviter d'insérer des phrases identiques dans le résumé (des phrases quasi-dupliquées ont été introduites par les organisateurs dans DUC 2005 et 2006) : une phrase est écartée si elle n'apporte pas de mots nouveaux au résumé. Cette analyse permet de déduire qu'une grande partie de la robustesse du

système provient des traitements annexes, et que la sélection de phrases en elle-même (MMR-LSA) est bénéfique pour des taux d'erreur inférieurs à 50%.

Au delà de la robustesse du système, l'observation du maintien des performances de MMR-LSA dans des conditions dégradées pose la question de la validité de la mesure Rouge. Cette mesure évalue la qualité du fond d'un résumé par son taux de rappel en *n*-grammes par rapport à un ensemble de résumés de référence. Bien qu'elle soit fortement corrélée avec les évaluations manuelles, les conditions dans lesquelles la mesure n'est plus représentative ne sont pas bien connues. Il serait intéressant de comparer dans les conditions d'un résumé « écouté », les performances Rouge d'une soumission fondée sur des données fortement dégradées et la perception par l'utilisateur de la qualité du contenu. Une autre piste serait de voir à quel point reproduire dans un résumé la distribution des mots dans les documents est une *baseline* performante. Cette dernière, bien que trop élaborée pour être considérée comme une *baseline*, pourrait bien nous amener à reconsidérer la notion de qualité dans les approches statistiques au résumé par extraction.

6.3 Conclusion

Nous avons prouvé dans ce chapitre que le système proposé est au niveau des systèmes état de l'art sur une tâche de résumé textuel. Pour cela, la méthode a été évaluée à travers une participation conjointe LIA-Thales à la campagne *Document Understanding Conference* (DUC) 2006. Cette soumission est une fusion de cinq systèmes de sélection de phrases (dont MMR-LSA, décrit dans ces travaux). En plus de cette évaluation ciblant le résumé textuel, nous avons dégradé les données DUC pour simuler les erreurs de la structuration automatique d'un contenu audio. Cette expérience montre que le système proposé est robuste à des erreurs uniformes (le type d'erreur le moins favorable pour un système de résumé) jusqu'à un taux d'erreur mots (WER) d'environ 40%. Les évaluations DUC ont tout-de-même montré que les approches par extraction aboutissaient généralement à une faible qualité de la structure des résumés. L'objectif du prochain chapitre est d'étudier des moyens de contourner cet aspect à l'aide d'interactions utilisateur complémentaires. Ce chapitre sera aussi l'occasion de mettre en valeur la chaîne de traitement complet « de l'audio à l'utilisateur », au sein du démonstrateur développé.