

---

# Une Méthode Contextuelle d'Extension de Requête avec des Groupements de Mots pour le Résumé Automatique

**Jean-François Pessiot, Young-Min Kim, Massih-Reza Amini  
Nicolas Usunier, Patrick Gallinari**

*Laboratoire d'Informatique de Paris 6  
104, Avenue du Président Kennedy  
75016 Paris, France*

*{pessiot,kim,amini,usunier,gallinari}@poleia.lip6.fr*

---

*RÉSUMÉ. Dans cet article nous décrivons les différentes étapes de construction du système de résumé extractif du LIP6 utilisé lors de la compétition Document Understanding Conferences (DUC2007). Ce système repose sur un module d'extension des mots de la question et du titre de chacun des sujets par des concepts de mots trouvés automatiquement avec un algorithme d'apprentissage non-supervisé. Cet algorithme est une version classifiante de l'algorithme EM. Chaque phrase des documents de la collection est ensuite caractérisée par un vecteur représentant les similarités de la phrase avec le titre, la question ainsi que leur version étendue. Le score final des phrases est alors trouvé en combinant manuellement ces similarités sur la base DUC 2006. Les résultats obtenus lors de cette compétition place le LIP6 respectivement 3<sup>ème</sup>, 2<sup>ème</sup> et 1<sup>ère</sup> suivant les trois mesures officielles de la compétition.*

*ABSTRACT. This paper describes the different steps which lead to the construction of the LIP6 extractive summarizer. The basic idea behind this system is to expand question and title keywords of each topic with their respective cluster terms. Term clusters are found by unsupervised learning using a classification variant of the well-known EM algorithm. Each sentence is then characterized by 4 features, each of which uses bag-of-words similarities between expanded topic title or questions and the current sentence. A final score of the sentences is found by manually tuning the weights of a linear combination of these features ; these weights are chosen in order to maximize the Rouge-2 AvF measure on the Duc 2006 corpus.*

*MOTS-CLÉS : Résumé automatique, Apprentissage non-supervisé, Extension de requêtes*

*KEYWORDS: Text Summarization, Unsupervised Learning, Query expansion*

---

## 1. Introduction

Le résumé est l'art de comprimer l'information. En RI, cette tâche avait initialement comme ambition de générer rapidement des résumés synthétiques à de nombreux documents. Cependant, les traitements linguistiques nécessaires pour trouver de tels résumés sont trop coûteux pour être appliqués à de grands corpus de documents (Sparck-Jones, 1993). Une simplification de la tâche largement étudiée consiste à extraire d'un document les passages (phrases ou paragraphes) les plus représentatifs de son contenu. L'extraction de ces passages est effectuée grâce à des heuristiques spécifiques, dont sept grandes classes ont été identifiées et énumérées dans (Paice *et al.*, 1993). Par exemple, une classe d'heuristiques consiste à comparer les mots d'un passage avec les mots du titre d'un document, une deuxième est de vérifier si le passage contient des marqueurs linguistiques (ou cue-words), comme *en conclusion*, *en résumé*, etc. Ces heuristiques allouent des scores réels aux passages, et les passages obtenant les meilleurs scores constituent le résumé automatique. Cette réduction (de l'abstraction à l'extraction) peut être vue comme la première étape vers la constitution d'un résumé synthétique ; pour résumer un texte les humains ont en effet tendance à surligner ses passages pertinents avant de le synthétiser.

Nous nous sommes placés dans le cadre du résumé extractif. L'état de l'art consiste dans ce cas à combiner les heuristiques d'un passage manuellement (Goldstein *et al.*, 1999) ou avec des méthodes génériques (Kupiec *et al.*, 1995). La solution adoptée par ces dernières est d'effectuer une combinaison numérique des heuristiques de façon automatique grâce à l'apprentissage statistique, en utilisant des bases de documents étiquetés : pour chaque document, les phrases<sup>1</sup> possèdent une étiquette 1 ou -1 correspondant respectivement au fait qu'elles font partie ou non d'un résumé de référence du document. Ces étiquettes sont déterminées soit manuellement soit, pour des cas particuliers et à des fins d'expérimentation, par des méthodes automatiques (Marcu, 1999) qui utilisent une information supplémentaire non disponible à l'algorithme d'apprentissage. Dans ces systèmes (Kupiec *et al.*, 1995, Amini, 2001), un algorithme de classification est alors utilisé pour déterminer la combinaison des scores des heuristiques qui a pour but d'associer à chaque phrase son étiquette 1 ou -1. (Marcu, 1999) propose en effet un algorithme d'alignement qui sur des articles scientifiques apparie l'abstract de ces articles aux phrases des articles et en extrait celles qui sont les plus ressemblant à l'abstract. Il montre ensuite sur différentes expériences que son approche est valide et que les phrases ainsi extraites sont très ressemblants (au sens d'un expert humain) aux vrais résumés extractifs des articles. Pour un nouveau document, le résumé est constitué des  $k$  phrases de ce document qui, d'après la fonction apprise, sont les plus susceptibles d'avoir pour étiquette 1, où  $k$  est un taux de compression et est généralement fixé à 10 ou 20% du nombre de phrases du document. Nous avons proposé un algorithme général d'ordonnancement dont l'objectif est d'apprendre un meilleur tri des phrases appartenant aux résumés de chaque docu-

---

1. Dans la suite nous choisirons les phrases comme unités d'extraction de base.

ment et nous avons montré que cet algorithme est plus performant que les algorithmes de classification proposés dans l'état de l'art (Amini *et al.*, 2007).

Dans cet article nous allons présenter une technique pour trouver des heuristiques pertinentes à base de clustering de mots. Le système de résumé que nous avons développé à base de ce concept s'est classé 1<sup>er</sup>, 2<sup>ème</sup> et 3<sup>ème</sup> parmi 32 participants sur les trois mesures officielles à la compétition DUC 2007 (Document Understanding Conference). La combinaison d'heuristiques dans ce cas s'est faite manuellement car comme nous allons l'aborder à la section suivante, les résumés dans cette compétition étaient à base de questions et nous ne disposions pas de résumés de références dédiées pour apprendre automatiquement cette combinaison.

Le plan de notre article est comme suit, dans la section 2 nous présenterons brièvement les compétitions DUC et la tâche à laquelle nous avons participé l'an dernier à DUC 2007. La section 3 présente la technique d'extension de requêtes à base de relation de co-occurrence que nous avons développée. Les sections 4 et 5 décrivent la technique de filtrage d'information non-pertinente et le scoring des phrases. Les résultats obtenus à la compétition sont donnés à la section 6.

## 2. Compétitions DUC

Les compétitions DUC ont été créées en Mars 2000 à l'initiative des agences DARPA (Defense Advanced Research Projects Agency), ARDA (Advanced Research and Development Activity) et NIST (National Institute of Standards and Technology). Ces agences travaillaient en parallèle sur des programmes proches de la thématique d'extraction de passages comme le programme TIDES (Translingual Information Detection Extraction and Summarization) de DARPA, *Advanced Question & Answering* de ARDA et *Text Retrieval Conferences* de NIST. Les compétitions DUC ont alors été mises en place dans le but de réunir les efforts consentis par ses différentes agences autour de la problématique du résumé de texte.

Ces compétitions sont maintenant financées par ARDA et sont tenues annuellement par NIST. Les premières compétitions étaient focalisées sur le résumé du contenu, des documents d'une collection donnée (appelé résumé générique mono-document). Depuis 2005, ces compétitions se sont intéressées au résumé d'un ensemble de documents par rapport à une question donnée (résumé multi-documents par rapport à une requête). Le but ici est de trouver la réponse à une question parmi un ensemble fixe de documents traitant du sujet de la question. Cette réponse ne doit pas comporter plus de 250 mots.

Une des difficultés à laquelle étaient confrontés les organisateurs de ces compétitions était l'évaluation des systèmes qui au début était basée sur le jugement humain pour déterminer de la *cohérence*, la *consistance*, la *lisibilité*, le *contenu* et la *grammaire* des résumés produits (Mani, 2001). Ainsi un effort humain de plus de 3000 heures était nécessaire pour évaluer chacun des systèmes participants à DUC 2003 (Over *et al.*, 2003). Beaucoup de travaux se sont alors intéressés à trouver des mé-

thodes d'évaluation automatiques. Une étude pionnière dans ce sens a été réalisée par (Lin *et al.*, 2003) qui ont montré que des méthodes similaires à celles utilisées en traduction automatique, comme les mesures ROUGE (Recall-Oriented Understudy for Gisting Evaluation) basées sur des comptages de n-grammes communs entre les résumés humains et machines, pouvaient être appliquées pour évaluer les résumés produits. Ainsi pour les deux dernières compétitions DUC, 3 évaluateurs humains fournissaient chacun un résumé synthétique à partir de 25 documents contenant une (ou des) réponse(s) à une question donnée. Pour chaque question, la sortie d'un système, d'une taille maximale de 250 mots, était alors alignée sur les 3 résumés humains et les performances ROUGE associées à chaque système étaient calculées en moyennant ces mesures sur l'ensemble des questions annotées.

Les mesures ROUGE sont des mesures de rappel et de précision sur des n-grammes présents dans les résumés produits. Avec le score manuel sur le contenu des résumés, les autres mesures officielles des compétitions DUC sont les performances ROUGE-2 et ROUGE-SU4.

Pour un résumé  $\mathcal{R}$ , la mesure rappel ROUGE-2 est calculée comme :

$$\text{Rappel ROUGE-2}(\mathcal{R}) = \frac{\sum_{\mathcal{R}_{ref} \in \{\mathbf{Rr}\}} \sum_{\text{gramm}_2 \in \mathcal{R}_{ref}} \delta(\text{gramm}_2, \mathcal{R}) \text{Nb}(\text{gramm}_2)}{\sum_{\mathcal{R}_{ref} \in \{\mathbf{Rr}\}} \sum_{\text{gramm}_2 \in \mathcal{R}_{ref}} \text{Nb}(\text{gramm}_2)}$$

Où,  $\mathbf{Rr}$  est l'ensemble des résumés de références,  $\mathcal{R}_{ref} \in \mathbf{Rr}$  est un résumé de référence contenant la réponse à la question posée,  $\text{Nb}(\text{gramm}_2)$  est le nombre total de bigrammes présents dans le résumé de référence et  $\delta(\text{gramm}_2, \mathcal{R})$  est le symbole de Kronecker prenant la valeur 1 si  $\text{gramm}_2$  est présent dans le résumé produit  $\mathcal{R}$  et 0 sinon.

La mesure précision ROUGE-2 est égale, quant à elle, à :

$$\text{Précision ROUGE-2}(\mathcal{R}) = \frac{\sum_{\mathcal{R}_{ref} \in \{\mathbf{Rr}\}} \sum_{\text{gramm}_2 \in \mathcal{R}_{ref}} \delta(\text{gramm}_2, \mathcal{R}) \text{Nb}(\text{gramm}_2)}{\sum_{\mathcal{R} \in \{\text{Résumés produits}\}} \sum_{\text{gramm}_2 \in \mathcal{R}} \text{Nb}(\text{gramm}_2)}$$

Sur la base de ces mesures, la F-mesure ROUGE-2 est calculée comme :

$$\text{Mesure-F ROUGE-2}(\mathcal{R}) = \frac{2 \times \text{Précision ROUGE-2}(\mathcal{R}) \times \text{Rappel ROUGE-2}(\mathcal{R})}{\text{Précision ROUGE-2}(\mathcal{R}) + \text{Rappel ROUGE-2}(\mathcal{R})}$$

Selon le même principe, d'autres mesures ROUGE-n peuvent être calculées.

La mesure ROUGE-S qui estime quant à elle le nombre moyen de paires de mots dans le résumé produit qui sont dans le même ordre que celles apparaissant dans les résumés de références. Par exemple si on considère les résumés produit et de référence suivants :

$\mathcal{R}_{ref}$  : *Le vélib a eu un franc succès.*

$\mathcal{R}$  : Le vélib a beaucoup de succès.

Les paires de mots apparaissant dans le même ordre dans  $\mathcal{R}_{ref}$  et  $\mathcal{R}$  sont : ('Le vélib', 'Le succès', 'vélib succès') et la mesure ROUGE-S de  $\mathcal{R}$  vaut  $\frac{3}{C_{|\mathcal{R}_{ref}|}^2}$ .

Cette mesure ignore toutefois les résumés qui contiennent des mots du résumé de référence mais pas de paires de mots communs. Par exemple la phrase : *Les organisateurs rencontrent leur premier succès franc avec vélib* a un score de ROUGE-S nul. Ainsi la mesure ROUGE-SU4 qui comptabilise des paires de mots communs et les mots apparaissant dans le résumé de référence, a été proposée pour pallier à ce problème.

### 3. Caractéristiques pour le résumé automatique

Dans cette section, nous décrivons les heuristiques de sélection des phrases. Chacune de ces heuristiques identifie des caractères particuliers des phrases qui tendent à apparaître dans celles qui doivent être extraites pour former un bon résumé, et, pour une phrase donnée, une heuristique renvoie un score réel, qui est d'autant plus grand que le caractère recherché a été identifié. Chaque phrase est alors décrite par un vecteur de scores fournis par les différentes heuristiques, où la valeur à une dimension donnée est le score renvoyé par l'heuristique correspondant à cette dimension. Le but de l'apprentissage est alors d'apprendre une combinaison linéaire de ces scores. L'étape préliminaire de la recherche d'heuristiques est donc primordiale pour la phase d'apprentissage. L'objectif est en effet d'avoir des caractéristiques indépendantes, chacune tenant compte d'un critère particulier de pertinence des phrases, puis de les combiner, afin d'obtenir une combinaison plus performante que la meilleure caractéristique.

(Paice *et al.*, 1993) regroupent les caractéristiques à prendre en compte pour le résumé automatique en sept catégories : 1) les marqueurs linguistiques (aussi appelés cue-words), 2) les acronymes, 3) les mots fréquents d'un document, 4) les mots-clefs du titre du document, 5) la position de la phrase dans le document, 6) la longueur de la phrase, et 7) les liens sémantiques entre les phrases. Ces caractéristiques ont été utilisées partiellement ou dans leur totalité dans (Kupiec *et al.*, 1995, Goldstein *et al.*, 1999). (Kupiec *et al.*, 1995) ont utilisé les marqueurs linguistiques, les acronymes, la similarité des phrases par rapport à une requête, la longueur des phrases ainsi que leur position dans le document. Les caractéristiques qui permettent de trouver les phrases pertinentes par rapport au contenu du texte (ou l'information souhaitée) sont issues de deux requêtes types. La première constituée des mots les plus fréquents de la collection de documents considérée, notée MFT (Most Frequent Terms) dans la suite, et la seconde est constituée des mots du titre du document considérée, notée title keywords dans la suite. Pour la compétition DUC, à la place des mots du titre, nous avons considéré les mots clés des questions. Dans les deux cas, mots du titre ou mots de la question il s'agit de trouver l'information pertinente par rapport à une source fixe ou une requête ouverte.

### 3.1. Extension de requêtes par lien sémantique ou local

Il a été montré que des enrichissements de la requête ouverte pouvaient améliorer les performances de façon très significative (Goldstein *et al.*, 1999). En effet, le titre du document ou une question donnée, ainsi que les phrases, contiennent peu de mots et sont donc sensibles aux variations linguistiques. Autrement dit, il faut pouvoir détecter dans les phrases d'un document les mots sémantiquement proches de ceux de l'information recherchée. Il est commun d'utiliser des techniques d'extension de requête (Goldstein *et al.*, 1999), soit en s'appuyant sur le thesaurus WordNet (Fellbaum, 1998) (lien sémantique), soit des techniques d'enrichissement de requêtes à base de Local Context Analysis (LCA) (Xu *et al.*, 1996).

### 3.2. Extension de requêtes par relation de co-occurrences

Nous avons proposé une nouvelle approche d'extension de requêtes à partir de groupement de mots qui, contrairement aux techniques locales comme le LCA qui ne considère que des co-occurrences locales au document considéré, permettent de prendre en compte les co-occurrences de mots dans le corpus de document tout entier. La création des groupements de mots est basée sur l'hypothèse  $\mathcal{H}_c$  que *deux mots co-occurrents dans le même contexte sont sémantiquement similaires*. Cette hypothèse s'interprète de la manière suivante : Les mots qui sont souvent utilisés ensemble dans un contexte (i.e. paragraphe) ont une forte probabilité de synonymie : pour décrire un phénomène on utilise souvent, dans un contexte local des synonymes relatifs au phénomène.

Pour former les groupes, nous reprenons la procédure décrite dans (Caillet *et al.*, 2004), qui consiste à représenter chaque mot  $w$  par un vecteur de sac-de-documents  $\vec{w} = \langle tf(w, d_i) \rangle_{i \in \{1, \dots, |D|\}}$  où  $D$  est la collection de documents et  $tf(w, d_i)$  est le nombre d'occurrences de  $w$  dans le document  $d_i \in D$ . Nous supposons que les termes sont générés indépendamment par un mélange de densités et que les composantes du mélange suivent la loi de Naïve-Bayes :

$$p(\vec{w} | \Theta) = \sum_{k=1}^c \pi_k p(\vec{w} | y = k, \theta_k)$$

Le regroupement des mots est ensuite effectué grâce à l'algorithme CEM (Caillet *et al.*, 2004), qui détermine les paramètres du modèle en optimisant la fonction de vraisemblance classifiante suivante :

$$L_{CML}(P, \Theta) = \sum_{w_j \in V} \sum_{k=1}^c \tilde{t}_{kj} \log p(\vec{w}_j, y = k, \Theta) \quad [1]$$

Où  $V$  désigne le vocabulaire,  $c$  est le nombre de classes et l'indice de classe  $k \in [1, c]$ . Les paramètres  $\Theta$  de ce modèle sont l'ensemble des probabilités de

|                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| D0614 - Quebec independence                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <p><b>Cluster contenant <i>quebec</i></b> : majority minister future prime chretien canadians federalist believe stay poll confederation unity center legislature uncertainty quebec province national face canada</p> <p><b>Cluster contenant <i>independence</i></b> : separatists united independence leaders need states public votes despite lucien create clear negotiations officials bouchard opposition france opinion independent</p> |
| D0705 - Basque separatism                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <p><b>Cluster contenant <i>basque and separatism</i></b> : basque separatist armed bilbao killed spain eta separatism independence police france batasuna nationalists herri</p>                                                                                                                                                                                                                                                                |

**Tableau 1.** Deux clusters de termes trouvés avec l’algorithme CEM à DUC 2006 et DUC 2007.

classes  $\pi_k = p(y = k)$  et les probabilités des documents appartenant aux clusters  $\{p_{ik}\}_{i \in \{1, \dots, |D|\}, k \in \{1, \dots, c\}}$ . Avec l’hypothèse d’indépendance de Naïve-Bayes les probabilités conditionnelles de classes s’expriment en fonction des paramètres du modèle comme  $p(\vec{w} | y = k) = \prod_{i=1}^{|D|} p_{ik}^{tf(w, d_i)}$ .

En dérivant  $L_{CML}$  par rapport aux paramètres  $\pi_k$  et  $p_{ik}$  et en utilisant les multiplicateurs de Lagrange pour renforcer les contraintes  $\sum_k \pi_k = 1$  et  $\forall k, \sum_i p_{ik} = 1$ , les estimées de  $\pi_k$  et  $p_{ik}$  au sens du maximum de vraisemblance s’écrivent :

$$\pi_k = \frac{\sum_{j=1}^{|V|} t_{kj}}{|V|} \quad p_{ik} = \frac{\sum_{j=1}^{|V|} t_{kj} \times tf(w_j, d_i)}{\sum_{j=1}^{|V|} \sum_{i=1}^{|D|} t_{kj} \times tf(w_j, d_i)}$$

Ce procédé permet donc de trouver des groupements de mots qui tendent à apparaître dans les mêmes documents. Le nombre de groupes à trouver est un hyperparamètre de l’algorithme. Nous l’avons fixé à  $\frac{|V|}{15}$  qui a conduit à de bonnes performances empiriques sur la base DUC 2006. Nous avons trouvé que les mots des groupes étaient aussi différents que ceux trouvés par les techniques utilisant les co-occurrences locales comme le LCA, et fournissent donc une information supplémentaire et indépendante de celle que nous avons déjà. À partir de ces groupements de mots nous avons enrichi les mots clés de la question et de son sujet en y ajoutant les mots contenus dans les mêmes clusters que ces mots clés.

Pour les deux dernières compétitions DUC, les 25 documents correspondant à chacun des sujets étaient tous pertinents par rapport à la question associée au sujet. Comme ces documents sont des dépêches d’agence de presse de taille assez courte, les mots faisant partie de la réponse co-occurrent souvent avec les mots clés de la question ou du sujet.

L’hypothèse  $\mathcal{H}_c$  selon laquelle nous avons construit les groupements de mots nous a ainsi permis de trouver partiellement les mots de la réponse. Le tableau 1 montre les

---

The Basque separatist group ETA conducted a weeklong unilateral ceasefire in Spain in June 1996. The truce ended in July with a series of bomb attacks on tourist resorts. In October 1997, ETA was preparing for another ceasefire when all 23 leaders of its political wing were jailed for distributing a pro-ETA video. In the first half of 1998 six slayings were attributed to ETA. Thousands demonstrated against the violence in the Basque region. The Socialist Party withdrew from the three-party Basque regional government. In August ETA said for the first time that it was opposed to street violence as a means of furthering the Basque cause. Basque nationalists joined with other political groups to urge ETA to seek a permanent end to the violence and in September ETA announced an open-ended ceasefire. The Spanish government agreed to hold peace talks, but said that there would be no discussion of Basque independence. ETA's Chief Hari Batasuna was arrested in France in August.

---

**Tableau 2.** *Un résumé synthétique pour la question Question(D0705).*

clusters de mots contenant les mots clés des sujets D0614 et D0705 des compétitions DUC 2006 et DUC 2007 trouvés par l'algorithme CEM. La question associée au sujet D0705 était :

Question(D0705) : *Describe developments in the Basque separatist movement 1996-2000.*

Il s'agissait de relater les faits sur les négociations entre les séparatistes ETA dirigés par Hari Batasuna et le gouvernement espagnol pour l'indépendance du pays Basque et la libération de leurs compagnons détenus en prison. Pendant la période des négociations, l'ETA et le gouvernement s'étaient mis d'accord sur une trêve qui s'est achevée après l'arrestation de Hari Batasuna en France. Un des résumés synthétiques fourni par les organisateurs pour cette question est montré dans le tableau 2. Sur cet exemple, 10 des 16 mots du cluster contenant les mots clés du sujet apparaissent 27 fois au total dans le résumé synthétique, alors que sur les deux mots clés du sujet il n'y a que *basque* qui est présent dans le résumé. Nous remarquons aussi que l'entité nommée *Hari Batasuna* qui demande une analyse linguistique poussée pour être identifiée comme étant liée au sujet est simplement détectée par notre algorithme en tant que termes co-occurent avec *Basque* et *separatism*.

D'une manière générale, en examinant la mesure rappel ROUGE-1 des requêtes *titre*, *titre étendu*, *question* et *question étendue* sur les deux compétitions DUC 2006 et DUC 2007, c'est à dire la masse que représente ces requêtes par rapport aux résumés humains, nous avons remarqué que les mots des requêtes étendues étaient proportionnellement plus présentes dans ces résumés que les requêtes non-étendues. Par exemple si on considère la requête *titre* pour la compétition DUC 2006, la valeur rappel ROUGE-1 maximale que peut atteindre une requête de taille moyenne égale à la taille moyenne de cette requête (3.46 mots) est de 0.014. La mesure rappel ROUGE-1 de la requête *titre* vaut quant à elle 0.0112 ce qui représente 80% de la valeur théorique. Ce pourcentage passe à 85% si on considère la requête *titre étendu*. En effet,

|                |                      | DUC 2006             | DUC 2007             |
|----------------|----------------------|----------------------|----------------------|
| # moy. de mots | Résumés synthétiques | 246.7                | 243.9                |
|                | Titre                | 3.46                 | 2.44                 |
|                | Titre étendu         | 21.8                 | 18.3                 |
|                | Question             | 10.04                | 6.93                 |
|                | Question étendue     | 61.12                | 59.4                 |
| Rouge-1        | Titre                | 0.0112 (80%)         | 0.0085 (85%)         |
|                | Titre étendu         | 0.075 ( <b>85%</b> ) | 0.068 ( <b>91%</b> ) |
|                | Question             | 0.0267 (65%)         | 0.019 (68%)          |
|                | Question étendue     | 0.177 ( <b>72%</b> ) | 0.185 ( <b>76%</b> ) |

**Tableau 3.** Statistiques sur l'effet de l'extension des requêtes

la taille moyenne et la mesure ROUGE-1 de la requête titre étendu sont respectivement de 21.8 et 0.075 pour cette compétition avec une valeur rappel ROUGE-1 maximale de 0.0883. Ces résultats sont récapitulés pour les deux compétitions dans le tableau 3. À l'aide de ces résultats nous en déduisons que, comme les requêtes étendues contiennent proportionnellement plus de mots des résumés cibles et que leur taille est 6 à 7 fois celles des requêtes de base, la similarité de ces requêtes avec les phrases des documents cibles devant faire parties des résumés a plus de chance d'être plus grande que la similarité de ces phrases avec les requêtes de base. Nous remarquons aussi que le gain de cette extension est plus accentué pour la compétition DUC 2007. Ainsi de la compétition DUC 2006 à DUC 2007, le gain en rappel ROUGE-1 de l'extension passe de +5% à +6% pour la requête titre et de +7% à +8% pour la requête question.

Avant le calcul des scores de phrases nous avons filtré les documents en enlevant les phrases les moins informatives par rapport aux questions posées. Pour cela nous avons appliqué l'algorithme d'alignement de (Marcu, 1999) qui extrait pour chaque document l'ensemble de ses phrases qui est le plus similaire à la question associée avec l'hypothèse sous-jacente que dans chaque document, le plus petit sous-ensemble des phrases qui contient la réponse à la question du sujet est aussi celui qui a la plus grande similarité sémantique avec la réponse. Cette algorithme est décrit à la section suivante.

#### 4. Algorithme d'alignement de Marcu

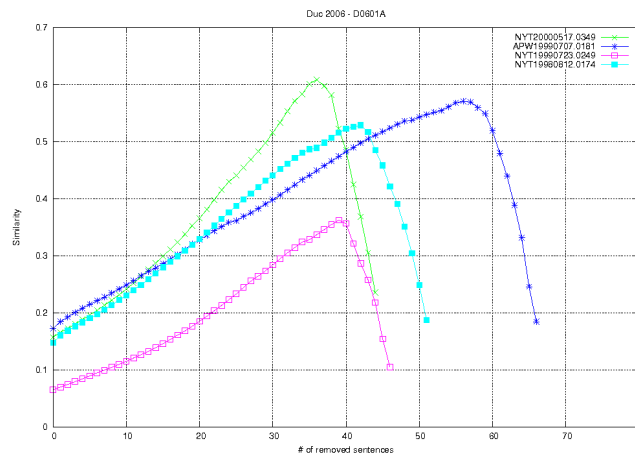
L'algorithme de Marcu (Marcu, 1999) calcule la similarité entre un ensemble de phrases  $\mathcal{S}$  et la question associée au sujet des documents contenant ces phrases,  $Q$  en utilisant la représentation sac-de-mots entre elles :

$$Sim(\mathcal{S}, Q) = \frac{\sum_{w \in \mathcal{S} \cap Q} c(w, \mathcal{S})c(w, Q)}{\sum_{w \in \mathcal{S}} c^2(w, \mathcal{S}) \sum_{w \in Q} c^2(w, Q)} \quad [2]$$

Où  $w \in Z$  (avec  $Z = \mathcal{S}$  ou  $Z = \mathcal{Q}$ ) signifie la présence du terme  $w$  dans  $Z$  et  $c(w, Z)$  est le poids associé à  $w$  dans  $Z$ . La pondération que nous avons choisie ici est :

$$c(w, Z) = tf(w, Z) \times \log(df(w))$$

Où  $tf(w, Z)$  est la fréquence de  $w$  dans  $Z$  et  $df(w)$  est le nombre de documents contenant  $w$ . Le choix de  $\log(df)$  à la place de  $\log(idf)^2$  qui est classiquement choisi en Recherche d'Information, s'explique par la construction des collections de DUC. Pour chaque sujet, les documents réunis pour le thème sont en effet tous pertinents par rapport au sujet et la question posée et les mots qui apparaissent fréquemment dans les documents de la collection relative à ce sujet sont ceux qui décrivent le mieux le thème.



**Figure 1.** Evolution de la mesure de similarité en fonction du nombre de phrases supprimées avec l'algorithme d'alignement de Marcu pour quelques documents de la collection D0601 de DUC2006.

L'algorithme de Marcu cherche itérativement à enlever la phrase dont la suppression augmente le plus, la similarité entre la question et l'ensemble restant des phrases (au sens de [2]). Cet algorithme s'arrête une fois que la suppression de n'importe quelle autre phrase de l'ensemble restant fait diminuer la similarité entre cet ensemble et la question. La figure 4 montre le comportement de cet algorithme sur quelques documents de la collection D601 de la compétition DUC 2006.

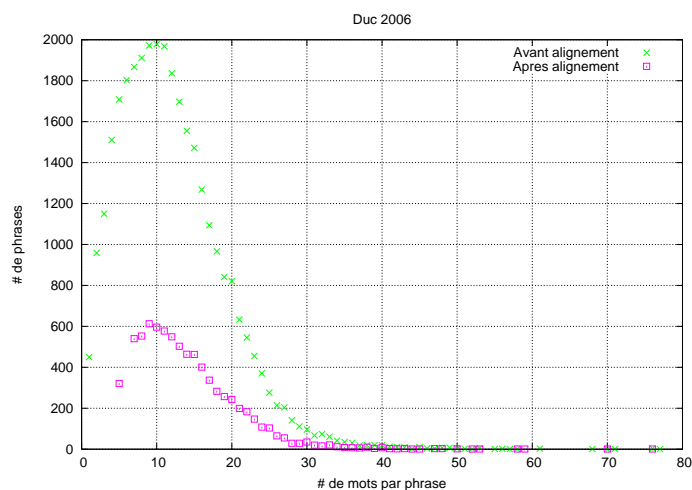
La figure 2 montre la distribution des mots dans les phrases avant et après l'application de l'algorithme de Marcu. Cette technique d'alignement ne change pas la

2.  $idf$  est l'inverse de  $df$  et il joue le rôle d'un terme de pénalisation pour les mots apparaissant fréquemment dans les documents d'une collection donnée.

| # moyen de        | Avant align <sup>t</sup> | Après align <sup>t</sup> |
|-------------------|--------------------------|--------------------------|
| phrases par sujet | 655.44                   | 156.7                    |
| mots par phrase   | 12.78                    | 9.78                     |
| Rappel ROUGE-1    | 0.96767                  | 0.89553                  |
| Précision ROUGE-1 | 0.01877                  | 0.06129                  |
| Rappel Rouge-2    | 0.56947                  | 0.44690                  |
| Précision Rouge-2 | 0.01073                  | 0.05299                  |

**Tableau 4.** Mesures ROUGE-1 et ROUGE-2 de l'ensemble des phrases présentes dans tous les documents pour chaque sujet avant et après la technique d'alignement de Marcu pour DUC 2006.

distribution des mots dans les phrases i.e. ce ne sont pas des phrases longues qui restent après l'alignement, ainsi pour la compétition DUC 2006 la longueur des phrases avant et après filtrage est normalement distribuée autour de 10 à 12 mots (figure 2).



**Figure 2.** La distribution de mots filtrés dans les phrases

En estimant les mesures ROUGE-1 et ROUGE-2 de toutes les phrases des documents avant et après alignement pour la compétition DUC 2006, nous avons remarqué que la précision de ces mesures sur la totalité des phrases après alignement a notablement augmenté alors qu'on ne note pas de grosse perte en rappel. Ces résultats montrent que proportionnellement il y a plus de phrases ne contenant pas l'information recherchée qui sont supprimées par l'algorithme de Marcu que de phrases pertinentes.

## 5. Mesures de similarité

Notre génération de caractéristiques est basée sur le calcul de similarité entre chaque phrase dans l'ensemble final des phrases et les requêtes à base du titre et de la question des sujets. Nous avons ainsi considéré trois requêtes :  $q_1$  représentant l'ensemble des mots clés de la question,  $q_2$  et  $q_3$  correspondant respectivement aux mots clés du titre et de la question ainsi que les mots des clusters contenant les mots des requêtes. Chaque caractéristique est alors décrite par  $f : \{requetes\} \times \{phrases\} \mapsto \mathbb{R}$ .

Où le score  $f$  d'une requête  $q$  avec une phrase  $p$  vaut  $f(q, s) = score(q, s)$ . Nous avons testé différentes fonctions de scores et avons trouvé sur la base DUC 2006 que les caractéristiques suivantes étaient les plus performantes :

| Caractéristique | Requête | Score                     |
|-----------------|---------|---------------------------|
| $F_1$           | $q_1$   | $termes\_communs(q_1, s)$ |
| $F_2$           | $q_1$   | $cosine(q_1, s)$          |
| $F_3$           | $q_2$   | $ldf(q_2, s)$             |
| $F_4$           | $q_3$   | $ldf(q_3, s)$             |

Où  $termes\_communs(q, s)$  est le nombre de mots communs entre la requête  $q$  et la phrase  $s$ ,  $cosine(q, s) = \frac{\sum_{w \in q \cap s} c(w, q)c(w, s)}{\sum_{w \in q} c^2(w, q) \sum_{w \in s} c^2(w, s)}$  où  $c(w, Z)$  est la même pondération de termes que celle utilisée avec l'algorithme de Marcu et  $ldf(q, Z) = \sum_{w \in q \cap Z} \log(df(w))$ .

| Caractéristiques | $F_1$ | $F_2$ | $F_3$ | $F_4$ |
|------------------|-------|-------|-------|-------|
| $F_1$            | *     | 0.198 | 0.186 | 0.141 |
| $F_2$            | *     | *     | 0.095 | 0.086 |
| $F_3$            | *     | *     | *     | 0.123 |

**Tableau 5.** Corrélation de Spearman entre différentes listes ordonnées obtenues avec les différentes caractéristiques.

Le tableau 5 montre la corrélation de spearman des caractéristiques que nous avons considérées. La corrélation de Spearman prend juste en compte l'ordre des phrases induites par ces mesures et non pas la valeur des rangs, les variations extrêmes dans ces valeurs n'interviennent donc pas dans le calcul de la corrélation. Des valeurs de corrélations basses suggèrent ici qu'il y a une faible relation linéaire entre les différentes listes ordonnées obtenues avec ces caractéristiques. La combinaison de ces scores permet ainsi de trouver plus de phrases pertinentes que chacune des caractéristiques séparément.

| DUC 2007 |                |                                     |                                     |
|----------|----------------|-------------------------------------|-------------------------------------|
| Id       | Moyenne        | Borne inf. de l'int. de conf. à 95% | Borne sup. de l'int. de conf. à 95% |
| D        | 0.17175        | 0.15322                             | 0.19127                             |
| C        | 0.14993        | 0.13372                             | 0.16741                             |
| J        | 0.14141        | 0.12265                             | 0.16274                             |
| G        | 0.13903        | 0.12312                             | 0.15385                             |
| E        | 0.13764        | 0.12413                             | 0.15315                             |
| B        | 0.13740        | 0.11372                             | 0.16061                             |
| F        | 0.13739        | 0.12097                             | 0.15530                             |
| A        | 0.13430        | 0.11765                             | 0.15108                             |
| I        | 0.13328        | 0.11017                             | 0.15481                             |
| H        | 0.12702        | 0.11448                             | 0.13995                             |
| 15       | 0.12285        | 0.11800                             | 0.12768                             |
| <b>4</b> | <b>0.11886</b> | <b>0.11467</b>                      | <b>0.12351</b>                      |
| 29       | 0.11725        | 0.11245                             | 0.12225                             |

**Tableau 6.** *Mesure-F ROUGE-2*

## 6. Résultats obtenus à la compétition DUC 2007

Comme nous ne disposons pas de phrases de résumés extraites à partir des documents cibles nous n'étions pas en mesure d'apprendre les poids de la combinaison. Nous avons montré dans (Amini *et al.*, 2007) que si une telle information est disponible il est possible d'apprendre à combiner ces caractéristiques en optimisant le rang moyen des phrases pertinentes au-dessus des phrases non-pertinentes. Pour la compétition DUC 2007 nous avons déterminé manuellement les poids de la combinaison pour lesquelles la mesure-F ROUGE-2 était optimale.

Pour chaque question, les phrases de tous les documents cibles associés étaient triées dans l'ordre décroissant des scores des phrases calculés après combinaison. Pour diminuer la redondance, sur les 10 phrases dans la tête de la liste, nous avons suivi (Conroy *et al.*, 2006) en éliminant celles qui avaient plus de 8 mots en communs avec les phrases mieux scorées qu'elles. Les résumés finaux étaient alors constitués en prenant les phrases restantes du début de la liste avec un nombre total de mots n'excédant pas 250.

Chacun des 45 résumés produits par les systèmes étaient évalués 3 fois par 3 juges humains différents (avec des identifiants allant de A à J), les scores ROUGE-2 et ROUGE-SU4 de chaque système étaient alors moyennés sur l'ensemble des scores donnés.

Les tableaux 6 et 7 donnent les résultats des mesures-F ROUGE-2 et ROUGE-SU4 des résumés produits par les humains et les trois premiers systèmes participants à la compétition. L'identifiant associé au système de résumé de LIP6 était 4. Notre

| DUC 2007 |                |                                     |                                     |
|----------|----------------|-------------------------------------|-------------------------------------|
| Id       | Moyenne        | Borne inf. de l'int. de conf. à 95% | Borne sup. de l'int. de conf. à 95% |
| D        | 0.21461        | 0.20154                             | 0.22922                             |
| C        | 0.19846        | 0.18350                             | 0.21478                             |
| J        | 0.19378        | 0.17834                             | 0.21139                             |
| E        | 0.19266        | 0.18147                             | 0.20490                             |
| F        | 0.19165        | 0.17905                             | 0.20506                             |
| A        | 0.18902        | 0.17749                             | 0.20182                             |
| G        | 0.18761        | 0.17638                             | 0.19886                             |
| B        | 0.18620        | 0.16685                             | 0.20543                             |
| H        | 0.18044        | 0.17067                             | 0.18967                             |
| I        | 0.18016        | 0.16292                             | 0.19648                             |
| 15       | 0.17470        | 0.16997                             | 0.17939                             |
| 24       | 0.17304        | 0.16800                             | 0.17769                             |
| <b>4</b> | <b>0.17007</b> | <b>0.16646</b>                      | <b>0.17381</b>                      |

**Tableau 7.** *Mesure-F ROUGE-SU4*

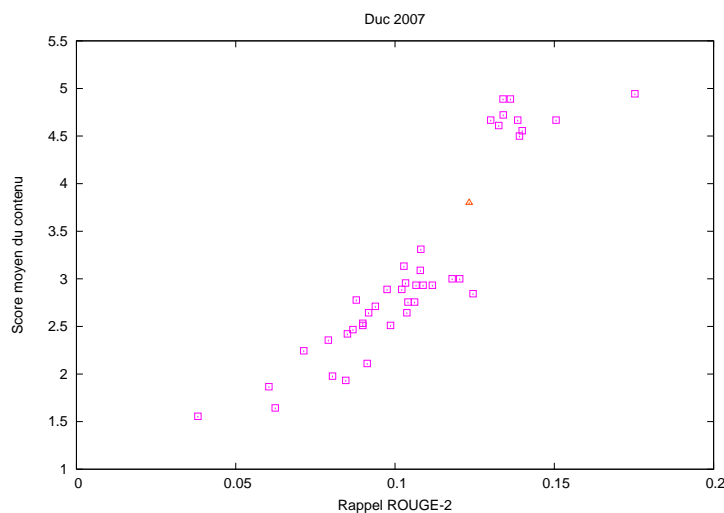
Le système s'est classé 2<sup>me</sup> et 3<sup>me</sup> respectivement suivant les mesures-F ROUGE-2 et ROUGE-SU4 à cette compétition.

Notre système a en outre réalisé le meilleur score linguistique parmi les systèmes participants à cette compétition (tableau 8). Ces scores étaient donnés par trois juges notant le contenu des résumés de 0 à 5.

| DUC 2007 |                        |
|----------|------------------------|
| Id       | Score Moyen du contenu |
| D        | 4.94                   |
| G        | 4.89                   |
| I        | 4.89                   |
| F        | 4.72                   |
| C        | 4.67                   |
| E        | 4.67                   |
| H        | 4.67                   |
| A        | 4.61                   |
| B        | 4.56                   |
| J        | 4.50                   |
| <b>4</b> | <b>3.80</b>            |
| 23       | 3.31                   |
| 14       | 3.13                   |

**Tableau 8.** *Scores linguistiques*

Le contenu des résumés était jugé par rapport aux réponses aux questions qu'ils contenaient. Ces résultats vont dans le sens de l'analyse donnée en section 3.2, ils renforcent l'idée que les mots des réponses co-occurrents avec les mots clés des questions et des titres. La figure 3 montre les scores moyens de contenu des systèmes participants en fonction de leur mesure Rappel ROUGE-2. Les deux groupements de points correspondent aux scores obtenus par les résumés humains (en haut à droite) et les scores des systèmes participants à la compétition (en bas à gauche). Sur cette figure, les scores de contenu (axe des ordonnées) séparent mieux, les résumés humains aux résumés produits par les systèmes, que la mesure Rappel ROUGE-2. Sur ce graphique, les scores de notre système sont montrés avec un triangle.



**Figure 3.** Les scores moyens du contenu des systèmes participants à DUC2007 en fonction de la mesure Rappel ROUGE-2

## 7. Conclusion

Nous avons présenté dans cet article une technique d'extension de requêtes à base d'une relation de cooccurrences de mots. Cette technique nous a permis d'avoir de bons résultats à la compétition de DUC2007 où il fallait trouver les réponses à 45 questions sur un ensemble de 25 documents cibles fournis avec une limitation de 250 mots pour les résumés produits. Ces documents étaient des dépêches de journaux et ils étaient tous pertinents par rapport à la question posée. Cette particularité a impliqué que les mots clés des questions co-occurrents avec une forte probabilité avec les mots de la réponse et l'hypothèse  $\mathcal{H}_c$  selon laquelle les mots co-occurrents dans le même contexte et avec la même fréquence sont thématiquement proches, nous a permis de trouver des groupements de mots dont ceux qui contenaient les mots de la question contenaient aussi certains mots de la réponse. Les scores entre les phrases et les re-

quêtes étendues étaient alors d'autant plus grands si ces phrases contenaient les mots de la réponse. Une perspective à ce travail consistera à la recherche d'une combinaison automatique des caractéristiques avec des modèles d'apprentissage (Usunier *et al.*, 2004).

## Remerciements

Ce travail a été sponsorisé en partie par le programme IST de la communauté Européenne, sous le réseau d'excellence de PASCAL, IST-2002-506778.

## 8. Bibliographie

- Amini M.-R., Apprentissage Automatique et Recherche d'Information : application à l'Extraction d'Information de surface et au Résumé de Texte, Thèse de doctorat, Université Pierre et Marie Curie, LIP6, Juillet, 2001.
- Amini M.-R., Tombros A., Usunier N., Lalmas M., « Learning Based Summarization of XML Documents », *Journal of Information Retrieval*, vol. 10, n° 3, p. 233-255, 2007.
- Caillet M., Pessiot J.-F., Amini M.-R., Gallinari P., « Unsupervised Learning with Term Clustering for Thematic Segmentation of Texts », *RIAO*, p. 648-656, 2004.
- Conroy J. M., Schlesinger J. D., O'leary D. P., Goldstein J., « Back to Basics : CLASSY 2006 », *Document Understanding Conference*, 2006.
- Fellbaum, *WordNet : An Electronic Lexical Database (Language, Speech, and Communication)*, The MIT Press, May, 1998.
- Goldstein J., Kantrowitz M., Mittal V. O., Carbonell J. G., « Summarizing Text Documents : Sentence Selection and Evaluation Metrics. », *SIGIR*, p. 121-128, 1999.
- Kupiec J., Pedersen J., Chen F., « A trainable document summarizer », *Proceedings of the 18<sup>th</sup> ACM SIGIR Conference*, p. 68-73, 1995.
- Lin C.-Y., Hovy E., « Automatic evaluation of summaries using N-gram co-occurrence statistics », *NAACL '03*, p. 71-78, 2003.
- Mani I., *Automatic Summarization*, John Benjamins Publishing Company, 2001.
- Marcu D., « The Automatic Construction of Large-Scale Corpora for Summarization Research », *Proceedings of the 22<sup>nd</sup> ACM SIGIR Conference*, p. 137-144, 1999.
- Over P., Yen J., « An Introduction to DUC 2003 : Intrinsic Evaluation of Generic News Text Summarization Systems », *Document Understanding Conference*, 2003.
- Paice C., Jones P., « The identification of important concepts in highly structured technical papers », *Proceedings of the 16<sup>th</sup> ACM SIGIR Conference*, p. 69-78, 1993.
- Sparck-Jones K., Discourse modeling for automatic summarizing, Technical report, Computer laboratory, university of Cambridge, 1993.
- Usunier N., Amini M.-R., Gallinari P., « Boosting Weak Ranking Functions to Enhance Passage Retrieval for Question Answering », *IR4QA-workshop, SIGIR*, 2004.
- Xu J., Croft W. B., « Query expansion using local and global document analysis », *Proceedings of the 19<sup>th</sup> ACM SIGIR Conference*, p. 4-11, 1996.