A chaque corpus sa méthode de découpage et une segmentation pour tous

Zohra Khalis, Caroline Tambellini, Loïc Maisonnasse

Laboratoire CLIPS IMAG – Université Joseph Fourier
385, rue de la bibliothèque - BP 53
38041 Grenoble cedex 9
{zohra.khalis, caroline.tambellini, loic.maisonnasse}@imag.fr

Résumé. Nous présentons ici les méthodes que nous utilisons dans le cadre de la campagne de fouilles de texte DEFT'06 dans le but de reconnaître automatiquement les segments thématiques de textes écrits en français provenant de différents domaines. Lors de cette campagne, nous évaluons des méthodes assez classiquement utilisées en segmentation thématique à savoir le TextTiling ou une méthode basée sur l'apprentissage mais aussi nous introduisons une nouvelle méthode basée sur la notion de cohérence de zones de texte que nous présentons dans cet article.

1 Introduction

DEFT'06 est la deuxième édition de la campagne d'évaluation du DEfi Fouille de Texte. L'objectif du défi cette année concerne la reconnaissance automatique des segments thématiques de textes écrits en français et provenant de différents domaines. La segmentation sert de base à différents traitements documentaires.

Elle permet d'isoler dans un document des zones répondant précisément à une requête thématique. Ceci est particulièrement utile dans un système de recherche d'informations. Elle permet aussi de sélectionner les différentes parties et les différents thèmes d'un texte lors de la création de résumés.

Pour DEFT'06, la segmentation s'appuie sur des corpus de documents de différents domaines écrits en français, à savoir des discours politiques, des textes juridiques et enfin un ouvrage scientifique. Le but du défi consiste donc à déterminer les segments thématiques de ces différents corpus. Nous présentons dans cet article les méthodes utilisées pour atteindre ce but. Après avoir présenté les méthodes de segmentations de l'état de l'art, nous présenterons les trois corpus et les éléments qui les rendent spécifiques et qui transforment la tâche de segmentation en trois tâches spécifiques aux corpus. Nous poursuivrons en présentant les méthodes mises en place pour gérer les spécificités de ces trois tâches.

2 La segmentation thématique

Beaucoup de moyens ont été imaginés pour segmenter un texte en thèmes cohérents. La principale différence entre ces méthodes tient au fait qu'elles soient supervisées ou non.

Les méthodes supervisées reposent sur une information à priori sur les classes importantes (leur nombre, leur signification, leurs caractéristiques statistiques), cette information est soit issue de bases de données, soit acquise lors d'une étape d'apprentissage. Parmi les méthodes supervisées, nous pouvons citer la méthode PLSA (Brants *et al.*, 2002) qui apprend les probabilités d'appartenance des termes à des classes sémantiques. D'autres méthodes se basent sur un apprentissage à base de modèles de Markov cachés (Amini *et al.*, 2000), ou encore proposent une classification des termes (Caillet *et al.*, 2004) (Chuang & Chien, 2004) et (Mekhaldi *et al.*, 2004).

Par opposition, les méthodes non supervisées qui ne nécessitent pas (ou très peu) d'information a priori, le caractère non supervisé porte sur l'estimation des caractéristiques statistiques des mots par des processus mathématiques de regroupement de données. L'interprétation des résultats obtenus n'est alors effectuée qu'à posteriori. Parmi les méthodes supervisées, nous pouvons citer la méthode TextTiling (Hearst, 97). Cette méthode mesure des similarités entre blocs adjacents en se basant sur la fréquence des mots et détermine les changements de thèmes par la détection de variations sur ces similarités. De même, la méthode de C99(Choi, 00) mesure la même similarité entre phrases mais de manière combinatoire sur l'ensemble des phrases texte. Les changements de thèmes sont déterminés par un algorithme de maximisation des valeurs de similarités trouvées. D'autres méthodes se basent sur les répétitions Dotplotting (Reynar, 98) et Segmenter (Kan, 98). Dans la première méthode, les répétitions sont répertoriées dans un graphe dont le nombre est minimisé afin de déterminer les changements de thèmes. Segmenter (Kan, 98) prend en compte les chaînes de répétitions auxquelles des poids sont associés en fonction de leur position dans le paragraphe et de leur catégorie syntaxique. La somme des poids des chaînes pour chaque paragraphe est calculée et détermine les changements de thèmes.

Il existe deux manières de segmenter, soit linéairement (les portions de texte trouvés sont adjacents), soit hiérarchiquement (on cherche à repérer les phrases correspondant à un même thème). Pour ce défi, la segmentation thématique utilisée est linéaire supervisée ou non.

3 Les corpus

3.1 Description des corpus

Le corpus est composé de trois corpus différents : discours politiques, textes juridiques et ouvrage scientifique. Le corpus de discours politiques est composé de discours politiques prononcés par des présidents de la république française (Valéry Giscard d'Estaing, François Mitterrand et Jacques Chirac). La segmentation thématique est basée sur la structure thématique des discours, les ruptures sont soit des changements de thème dans le discours soit des changements de discours.

Le corpus de textes juridiques est composé d'articles de lois de l'Union Européenne. Les segments thématiques sont les articles des lois. Il faut alors détecter les articles qui traitent du même sujet. Le corpus ouvrage scientifique est composé, quant à lui, du livre "Apprentissage Artificiel" d'Antoine Cornuéjols et Laurent Miclet (éditions Eyrolles). Avec ce corpus, les segments thématiques à retrouver sont les différentes sections (chapitres, sections, sous-sections, sous-sous-sections). Pour ce corpus, les titres des différentes sections ainsi que les figures, tableaux et les équations ont été supprimés. Le but est de déterminer la première phrase de chaque section.

3.2 Etude des corpus

Afin de définir la méthode la plus adaptée à ce défi, nous avons étudié les différents corpus pour déterminer leurs caractéristiques, et les traitements à effectuer.

3.2.1 Corpus politique

Le corpus politique regroupe les discours prononcés par des présidents de la république française. Sur ce corpus nous constatons que les ruptures ont un vocabulaire remarquable, par exemple un certain nombre vont contenir des mots tel que « madame, monsieur » représentatifs des phrases qui introduisent les discours. Les discours ayant une structure forte, les mots de liaison vont aussi être fortement discriminants pour déterminer les phrases de rupture, nous remarquons par exemple l'utilisation de mots tel que « enfin » « ensuite ». Sur ce corpus nous avons donc exploré un apprentissage du vocabulaire discriminant les phrases de rupture.

3.2.2 Corpus juridique

Le corpus juridique a la particularité d'apporter un indice supplémentaire au niveau de la segmentation. En effet, tous les segments thématiques commencent par la même phrase puisqu'un segment thématique débute toujours par un nouvel article de lois. De ce fait, on sait que forcément un changement sera de la forme « Article ». Compte tenu de cette constatation, une méthode basée sur une mesure de similarité entre blocs est particulièrement adaptée puisque nous pouvons définir aisément des blocs. Une adaptation de la méthode du TextTiling a donc été choisie.

3.2.3 Corpus scientifique

La particularité principale de ce corpus est la présence récurrente de digressions. En effet pour illustrer ces propos, l'auteur donne des exemples basés sur les différentes espèces d'oiseaux (anitidés). La méthode de cohérence est particulièrement adaptée, en effet lors de l'apparition de ces digressions la similarité varie et provoque des ruptures de thème à tort. L'évaluation de la cohérence permet de résoudre ce problème, c'est pourquoi nous utiliserons cette méthode dans ce contexte.

3.3 Prétraitement du corpus

Le but de la segmentation est d'extraire des corpus des segments de thème, nous nous intéressons donc au vocabulaire porteur de sens contenu dans les segments. Nous choisissons d'utiliser la forme lemmatisée des mots contenus dans les documents. L'utilisation des lemmes a l'avantage de regrouper sous une seule forme l'ensemble des flexions d'un mot. Ces lemmes sont obtenus par l'utilisation d'un analyseur morphosyntaxique TreeTagger[1] sur les phrases. Cet analyseur nous fournit pour chaque mot de la phrase originale, le lemme et la catégorie grammaticale correspondante à ce lemme. Les lemmes extraits ne sont pas tous représentatifs du thème ; les articles, les prépositions ne permettent pas de définir un thème.

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

A chaque corpus son découpage et une segmentation pour tous

Les méthodes que nous proposons sélectionnent donc un certain nombre des catégories grammaticales extraites par TreeTagger pour ne conserver que les lemmes les plus pertinents pour la tâche.

4 Méthode du TextTiling

La méthode du TextTiling (Hearst 1997) recherche les ruptures de thèmes et les identifie lorsqu'un bloc du document présente un moins grand nombre de mots traitant du thème. La méthode du TextTiling (cf. figure1) découpe tout d'abord (1) le document en blocs composés d'un nombre fixe de phrases (3 à 5 phrases généralement). Ensuite, (2) toutes les paires des blocs adjacents de textes sont comparées et une valeur de similarité leur est attribuée. (3) La suite résultante des valeurs de similarités, après être mise sous forme de graphes et aplanie, est examinée pour déterminer les pics et les vallées sur le graphique. (4) Des valeurs de similarités élevées, impliquant que les blocs adjacents se suivent de façon logique, sont susceptibles de former des pics, tandis que des valeurs de similarités faibles, indiquant une potentielle limite entre les blocs, créent des vallées. Un pic correspond donc à deux blocs fortement liés thématiquement alors qu'une vallée correspond à une rupture de thèmes. Chaque vallée est donc considérée comme une rupture de thèmes et correspond à une limite entre deux blocs thématiquement différents.

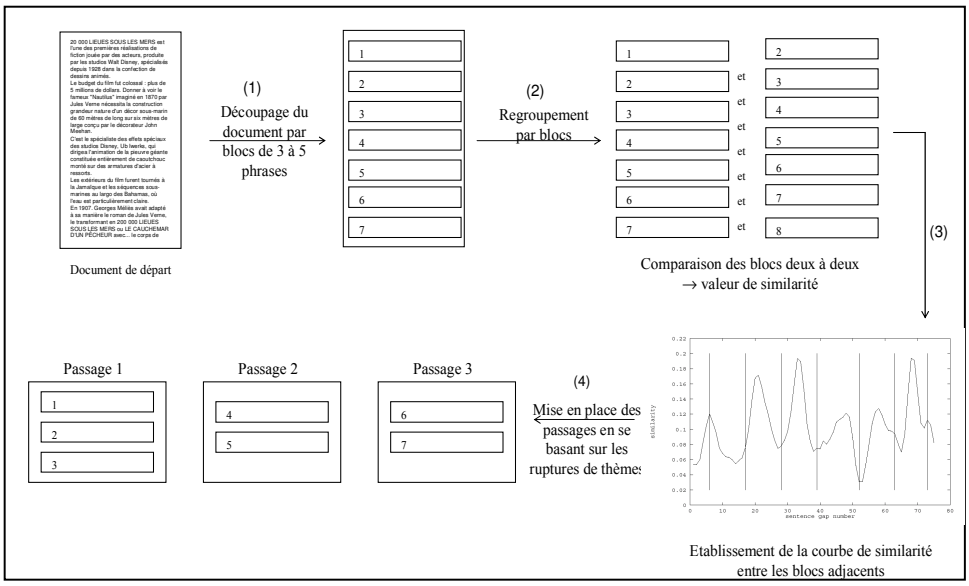


FIG. 1 – *Méthode du TextTiling*

4.1 Adaptation au contexte DEFT'06

Afin de déterminer les changements thématiques, nous implémentons une adaptation de la méthode du TextTiling. Nous gardons le principe du TextTiling et nous l'adaptons au contexte des textes juridiques. Nous ne gardons pas toutes les catégories morphosyntaxiques

de mots. Ainsi seuls, les verbes, les adjectifs, les noms communs et les noms propres. Les principales étapes du processus sont :

Découpage du document en blocs (1) ;
Calcul de similarité des blocs pris 2 à 2 (2) ;
Détermination du seuil permettant d'identifier les ruptures thématiques (3) ;
Détermination des ruptures (4) ;

La particularité dans l'utilisation du TextTiling ici et que nous n'utilisons pas des blocs de taille fixe mais des blocs que nous formons à partir des indices du document, la « phrase » « Article ». Ainsi, un bloc correspond à l'ensemble des phrases comprises entre 2 « phrases » « Article ». Une fois ces blocs de phrases formés (1), la similarité entre les blocs pris deux à deux est calculée (2) :

$$sim(a,b) = \frac{\sum_{t=1}^{n} w_{t,a} w_{t,b}}{\sqrt{\sum_{t=1}^{n} w_{t,a}^2 \sum_{t=1}^{n} w_{t,b}^2}}$$

où t varie pour tous les termes du document et $w_{t,a}$ est le poids tf.idf[2] assigné au terme t dans le bloc a.

Une fois cette similarité calculée, il faut fixer le seuil qui permettra d'identifier les ruptures. Après quelques tests, la valeur 0 donne les meilleurs résultats et cette valeur est donc prise comme seuil (3). Enfin, pour chaque valeur inférieure au seuil, on considère qu'il existe une rupture thématique entre les deux blocs correspondant à cette valeur de similarité (4). Si la valeur de similarité entre un bloc A et un bloc B est inférieure au seuil, la première phrase du bloc B est renvoyée.

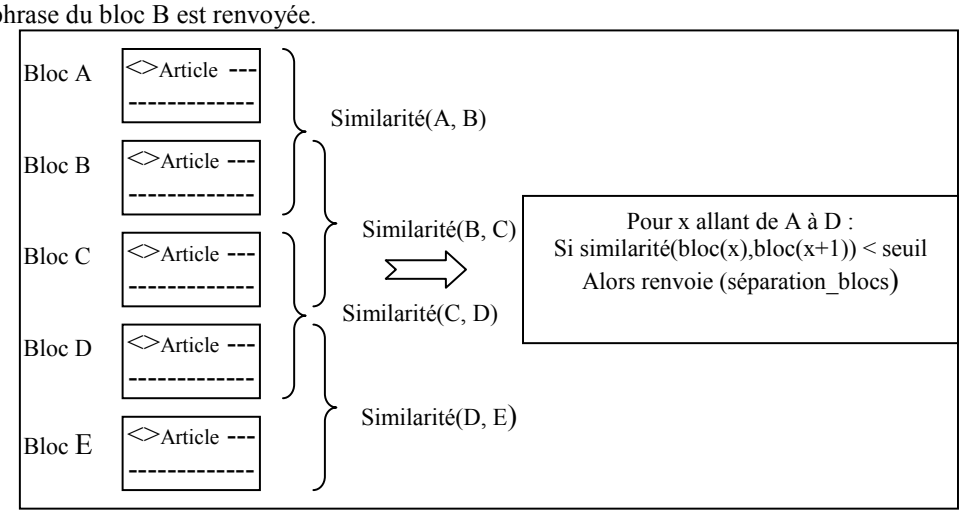
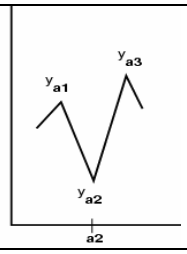


FIG. 2 – Principe de détermination des ruptures

[2] $\mathbf{tf.idf} = \frac{\text{nombre d'apparitions du terme dans le bloc}}{\text{nombre de blocs contenant le terme}}$

A chaque corpus son découpage et une segmentation pour tous



La méthode de TextTiling considère qu'il y a une vallée lorsque $(y_{a1} - y_{a2}) + (y_{a3} - y_{a2}) > \delta$, δ étant un seuil à fixer (cf. figure 3). Toutefois, nos expérimentations ont montré que l'utilisation ou non de cette contrainte donnait des résultats sensiblement égaux.

FIG. 3 – Principe des vallées du TextTiling

4.2 Limites de la méthode

Cette méthode connaît toutefois certaines limites. En effet, en utilisant des blocs formés entre deux phrases « Article », il existe un déséquilibre au niveau de la taille des blocs. Ainsi, on peut avoir des blocs comprenant seulement une ou deux phrases. Ceci pose un problème au niveau du calcul de similarité entre deux blocs, car un bloc court aura tendance à être considéré comme une rupture à tort. La méthode de TextTiling tente de tenir compte de cela en vérifiant la validité des vallées (cf. figure 3) et en utilisant pour le calcul de similarité la mesure cosinus.

5 Apprentissage des phrases de rupture

Sur le corpus politique les phrases de rupture ont souvent un vocabulaire typique. Par conséquent nous nous intéressons à la détection de ces phrases par un apprentissage. A partir des lemmes extraits par TreeTagger, et suite à une sélection des catégories grammaticales (nom, verbe, adjectif, pronom), chaque phrase est représentée sous la forme d'un vecteur de lemmes, ces vecteurs servent de base aux différents apprentissages. Nous divisons le corpus d'entraînement en deux parties l'une destinée à l'apprentissage, l'autre destinée à l'évaluation de l'apprentissage. Les vecteurs correspondant aux phrases de la partie apprentissage sont séparés en deux ensembles, d'une part ceux correspondant aux ruptures et d'autre part ceux correspondant au reste du texte. L'apprentissage sur ces deux ensembles nous permet d'établir le profil des phrases de rupture et le profil des phrases de non rupture. Ces *profils* sont représentés sous la forme de vecteurs de termes où chaque terme est associé à un poids représentant sa capacité à distinguer le profil. Nous comparons ensuite une à une les phrases de la partie évaluation de notre corpus avec chacun des profils en effectuant le produit scalaire du vecteur de la phrase et du profil, ce qui fournit un score de correspondance au profil pour chaque phrase. Enfin nous sélectionnons pour chaque phrase le profil qui lui correspond le mieux, c'est à dire celui ayant le meilleur score.

5.1 Formules d'apprentissage

Deux formules d'apprentissage sont évaluées, l'une basée sur la formule de *Rocchio,* l'autre sur la formule utilisée dans (Brouard, 2002) (*N*). Ces deux formules sont présentées ci-dessous :

Rocchio	$w_{i,j} = \alpha \frac{\lvert Qi \cap Pj \rvert}{\lvert P_j \rvert} - \beta \frac{\lvert Qi \cap \overline{Pj} \rvert}{\lvert \overline{P_j} \rvert}$, $\alpha = \beta = 1$
N	$w_{i,j} = \frac{\lvert Q_i \cap P_j \rvert}{\lvert P_j \rvert} * \frac{\lvert Q_i \cap P_j \rvert}{\lvert Q_i \rvert}$

TAB. 1 – Pondération pour l'apprentissage des phrases de rupture

Où : P_j : ensemble des phrases de type j, j étant une rupture ou une non rupture
Q_j : ensemble des phrases contenants le lemme i

5.1.1 Evaluation

Nous évaluons notre méthode d'apprentissage sur le corpus politique et sur le corpus scientifique. Aucun apprentissage n'est effectué sur le corpus juridique, en effet il n'est pas possible d'apporter une information supplémentaire sur ce corpus à l'aide d'un apprentissage, chaque rupture ayant la forme « article X ». Nous avons découpé en deux parties les corpus d'entraînements sur lesquels nous évaluons notre méthode. Pour le corpus politique, les phrases 1 à 150000 sont utilisées pour l'apprentissage et de 150001 à 303373 pour l'évaluation. Sur le corpus scientifique, l'apprentissage est effectué sur les phrases de 1 à 1999 et l'évaluation sur les phrases de 2000 à 4722. Pour chacun des corpus nous évaluons les deux formules d'apprentissage à l'aide du Fscore simple, proposé par les organisateurs de DEFT 06, les résultats sont décrits dans le tableau 2.

	Politique	Scientifique
Rocchio	**0,2013**	**0,1816**
N	0,1731	0,0097

TAB. 2 – Fscore résultat des fonctions d'apprentissage

Sur les deux corpus la formule de Rocchio donne les meilleurs résultats. Sur le corpus scientifique la formule N donne des résultats faibles, cela provient du fait que seulement 2 phrases sont considérées comme appartenant au profil de rupture. Cette formule est plus sélective que la méthode de Rocchio. Le corpus scientifique n'y est pas adapté du fait de la petite quantité des données d'apprentissage.

5.1.2 Variation du seuil sur Rocchio

Nous avons par la suite choisi d'utiliser uniquement la formule de Rocchio. Pour cette formule il n'est pas nécessaire d'établir les deux profils, l'un étant l'opposé de l'autre. Nous calculons donc uniquement le profil des ruptures. Le résultat fourni donne deux classes, la première contenant les phrases ayant un score supérieur à un seuil et appartenant à la classe des ruptures et l'autre contenant les phrases ayant un score inférieur à ce même seuil et par conséquent n'étant pas des ruptures. Les résultats précédents correspondent à l'utilisation d'un seuil de 0, ce seuil étant peu sélectif, nous évaluons ici plusieurs seuils supérieurs. L'ensemble des seuils évalués est présenté sur le tableau 3.

A chaque corpus son découpage et une segmentation pour tous

		politique			scientifique		
		Fscore	Fscore2	Fscore3	Fscore	Fscore2	Fscore3
seuil	0	0,2013	0,3958	0,5434	0,1816	0,4463	**0,6663**
	0,05	0,2299	**0,4251**	**0,5602**	0,2126	**0,4573**	0,6643
	0,1	0,2382	0,4185	0,5321	0,2231	0,4499	0,6314
	0,125	0,2537	0,3921	0,4978	**0,2307**	0,4483	0,6355
	0,15	**0,26**	0,3882	0,4823	0,2225	0,4325	0,6050
	0,175	0,2587	0,3759	0,4629	0,2036	0,3842	0,5386

TAB. 3 – Résultats en Fscore de la variation du seuil

Les résultats montrent l'importance du seuil, en effet la sélection du meilleur seuil permet d'obtenir une amélioration supérieure à 25% par rapport au seuil de base. Nous remarquons cependant que les seuils ne sont pas les mêmes en fonction des corpus. Cela provient du fait que dans le corpus politique il existe des ruptures fortement identifiables qui par conséquent ont un score élevé, telles que celles introduisant les discours, alors que celles-ci ne se retrouvent pas dans le corpus scientifique.

6 La méthode de cohérence

Cette méthode est particulièrement adaptée au corpus scientifique car la présence récurrente de digressions pour illustrer les propos basés sur les différentes espèces d'oiseaux (anitidés) provoque des ruptures de thème. La mesure de cohérence permet de les englober et de ne pas trop en tenir compte lors de la détermination des changements de thèmes.

Cette méthode est basée sur le fait qu'à l'intérieur d'un segment thématique les phrases doivent être fortement liées entre elles et peu liées avec les phrases n'appartenant pas au dit segment. Le but est donc de détecter les segments sous cette forme. Une phrase est intégrée à l'intérieur d'un segment thématique si elle et les phrases qui l'entourent sont fortement liées.

Cette méthode repose sur la mesure de l'intégration d'une phrase à son contexte local. Le contexte local est défini par une zone délimitée. L'intégration d'une phrase à ce contexte se mesure par le nombre de liens qu'elle entretient avec les autres phrases de la zone ainsi que le nombre de liens reliant une phrase la précédant à une phrase la succédant. Un lien est établi entre deux phrases quand la similarité entre elles est supérieure à un seuil.

Le calcul de la cohérence s'effectue en passant par une matrice de similarité puis de lien et enfin de cohérence. La dernière étape consiste en l'extraction des frontières des segments thématiques.

6.1 Le découpage physique

Le découpage physique découpe le document initial par phrases en ne prenant en compte que les noms, les verbes et les adjectifs ; les mots composés, quant à eux, ne sont pas pris en compte. L'algorithme prend en entrée une liste de phrases composées de lemmes. Chaque phrase est ensuite représentée dans l'espace vectoriel des lemmes à l'aide d'un vecteur dont les dimensions sont pondérées par la fréquence d'apparition du lemme dans la phrase (cf. figure 4).

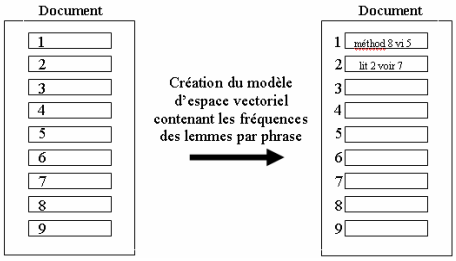


FIG. 4 – Découpage du texte en phrase et calculs des fréquences

6.2 Calcul de la similarité : génération d'une matrice de cohérence

Trois étapes sont nécessaires à l'établissement de la matrice de cohérence. Tout d'abord, un calcul de similarité entre les phrases de la zone délimitée. Ensuite, la binarisation de la matrice. Enfin l'établissement de la matrice de cohérence.

Pour commencer, nous établissons une zone de comparaison (cf. figure 5) qui correspond à une méthode intermédiaire entre les zones adjacentes (a) et toutes les zones (b). Elle permet d'avoir un contexte en amont et en aval mais n'englobe pas tout. Elle se focalise sur les éléments les plus pertinents qui sont les éléments proches et elle ne prend pas en compte des éléments trop éloignés qui n'appartiennent pas au segment thématique étudié. L'utilisation d'une zone intermédiaire (c) ne se fait pas uniquement pour limiter les calculs mais pour mieux cerner l'espace autour duquel le thème d'une phrase peut se propager.

(a) zones adjacentes (b) toutes les zones (c) zones dans un espace intermédiaire

FIG. 5 – Représentation d'une matrice et d'une zone matricielle

6.2.1 Calcul de la matrice de similarité

Une fois la zone établie, nous calculons l'ensemble des similarités entre les phrases au sein de cette zone. On calcule la matrice de similarité M, en utilisant pour toute phrase i une zone de comparaison. La zone de comparaison de cette phrase i est définie par l'ensemble des phrases comprises entre $i - \delta$ et $i + \delta$ où δ est la taille de la zone de comparaison.

Les similarités entre phrases sont ensuite calculées en utilisant la mesure du cosinus, selon la formule suivante :

$$M(i,j) = \frac{\sum_j f_{i,y} \times f_{j,y}}{\sqrt{\sum_i f_{i,y}^2 \times \sum_j f_{j,y}^2}} \quad \text{si } j \in [i-\delta, i+\delta]$$

$$M(i,j) = 0 \quad \text{sinon}$$

Où : $f_{x,y}$ la fréquence du mot x dans la phrase y .

Cette mesure est appliquée pour toutes les paires de phrases comprises dans l'intervalle afin de générer la zone de comparaison de similarité c'est à dire la matrice de similarité M(i,j). Nous pouvons supposer que si deux phrases sont fortement similaires cela signifie qu'elles sont à l'intérieur d'un même segment thématique et que de même, si une phrase se situe entre des phrases fortement similaires, cette phrase est à l'intérieur du segment thématique. Il est donc intéressant de prendre en compte ces phénomènes dans le processus de segmentation. En effet, dans aucune des méthodes existantes l'accent n'est mis sur la notion de liens et de contexte d'une phrase. Par conséquent, nous proposons une détection de ruptures alternative basée sur la détection de liens par la mesure, à l'aide d'un cosinus, de la similarité, et sur l'évaluation, à l'aide d'un comptage des liens entre phrases, de l'intégration d'une phrase à son contexte.

Pour atteindre notre but, à partir de la matrice de similarité, une sélection des similarités est effectuée et une matrice de cohérence est établie.

6.2.2 Calcul de la matrice de cohérence

La matrice de cohérence notée MC est la mesure des liens entre les phrases. Pour cela, chaque valeur de la matrice supérieure à un seuil va induire que les deux phrases paramètres de cette valeur sont incluses dans un même segment thématique. Ce seuil de similarité noté γ va permettre de déterminer si deux phrases sont considérées comme appartenant au même segment thématique.

Par exemple, si *M(phr1,phr10)* > *γ* cela signifie que les phrases allant de 1 à 10 appartiennent à un même segment thématique. Ceci va être traduit par un lien entre ses phrases. On définit :

$$Lien(i,j)) = \begin{cases} 1 \text{ si } M(i,j) > \gamma \\ 0 \text{ sinon} \end{cases}$$

L'intégration de la phrase *x* est évaluée par le nombre de liens *L* entre *y* et z tel que *y* inférieur ou égal à *x* et *z* supérieur ou égal à *x*. Une fois les liens définis, on peut établir la matrice de cohérence. On définit la matrice de cohérence comme :

$$MC(i,j) = \sum_{k=i-\delta}^{i} \sum_{l=j}^{j+\delta} Lien(k,l)$$

La courbe de cohérence correspond au nombre de liens sur la diagonale de la matrice de cohérence ce qui représente l'intégration d'une phrase à son contexte. La diagonale a donc de l'importance car elle représente l'ensemble des liens reliant la phrase au segment thématique.

6.3 Extraction des zones thématiques

A partir de la diagonale de la matrice, une courbe de cohérence est construite (cf. figure 6). Lorsque la courbe est forte, le nombre de liens encadrant la phrase en cours est fort, la phrase est alors fortement intégrée entre les phrases qui l'entourent, elle s'intègre donc dans un segment long. Deux phases sont donc à remarquer, si la courbe croît, les phrases sont de plus en plus liées entre elles alors qu'elles l'étaient peu avec les phrases précédentes : on commence donc un nouveau passage thématique. Au contraire si la courbe décroît alors les phrases sont de moins en moins liées (notamment avec les phrases suivantes) donc le thème est en train de changer. À partir de cette courbe, les ruptures de thème sont extraites, elles correspondent aux minimums locaux.

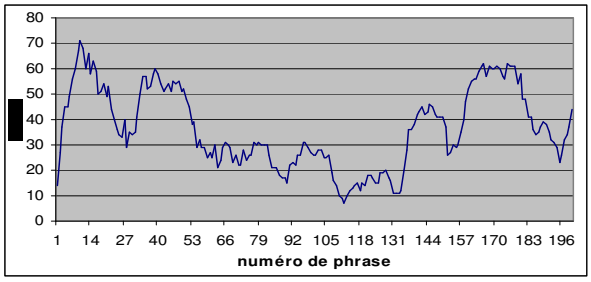


FIG. 6 – Courbe de cohérence d'un texte politique de 200 phrases

Une combinaison avec la méthode d'apprentissage de Rocchio est réalisée. Les valeurs données à chaque phrase lors de l'application de la méthode de Rocchio permettent de filtrer les minimums locaux de la courbe de cohérence. En effet, si la méthode de Rocchio donne une valeur négative à la phrase qui est considérée comme un minimum alors elle n'est pas considérée comme un changement thématique.

6.4 Evaluation des paramètres utilisés par la méthode

6.4.1 Choix de la taille de la zone de comparaison

Il faut définir la taille de la zone de comparaison. Celle-ci ne doit pas être trop petite, sinon les segments thématiques ne sont pas compris dans cette zone. Elle ne doit pas être non plus trop grande car elle prendrait en compte des éléments ne pouvant pas faire partie du même segment thématique lors de la segmentation. Le tableau suivant représente le Fscore sur le corpus scientifique pour l'implémentation de la méthode proposée en faisant varier la taille de la zone de comparaison de 5 à 100.

De plus, les meilleurs résultats, en ne tenant compte que de la taille de la zone de comparaison, sont ceux dont la taille de la zone est inférieure ou égale à la moyenne de la taille des segments (cf. tableau 4). Cela montre que plus la taille est grande, plus les résultats sont mauvais ce qui confirme l'hypothèse que plus la zone de comparaison est grande, plus elle a de chance de comparer des éléments qui n'ont aucun rapport entre eux et par conséquent créer des liens qui correspondent à des rappels de thème (digressions).

A chaque corpus son découpage et une segmentation pour tous

Taille de la zone	Moyenne des résultats
5	0,191
10	0,191
15	0,188
20	0,182
25	0,172
30	0,167
40	0,164

TAB. 4 – résultat de la variation de la taille des segments sur le corpus scientifique

6.4.2 Choix du seuil de similarité lors de l'établissement de la matrice de lien

Il faut déterminer le seuil de similarité à partir duquel deux phrases sont considérées comme appartenant au même segment thématique.

Le tableau 5 représente le Fscore sur le corpus scientifique pour l'implémentation de la méthode proposée en faisant varier le seuil de similarité. Les résultats montrent que le Fscore est meilleur lorsque la similarité entre deux phrases est supérieure à 0,5 (cf. tableau 5). La similarité est comprise entre 0 et 1, 0,5 est donc la valeur limite pour considérer que deux phrases se ressemblent ou non. Ceci confirme bien que les liens à prendre en compte doivent vraiment refléter le fait que les phrases sont similaires. En effet, si le seuil est placé en dessous de 0,5 les liens prennent en compte des phrases qui sont similaires mais pas suffisamment pour dire qu'elles font parties du même segment. De même, si le seuil est fixé à une valeur supérieure à 0,5 certains liens exprimant l'appartenance des phrases à un même segment sont perdus et par conséquent, la segmentation est biaisée.

Seuil	>0,3	> 0,4	> 0,5	> 0,6	> 0,7
Moyenne	0,164	0,161	0,172	0,170	0,164

TAB. 5 – variation du seuil de similarité sur le corpus scientifique

6.5 Conclusion

La méthode ainsi construite semble combler les lacunes des méthodes existantes qui prennent en compte soit seulement les phrases adjacentes soit l'ensemble des phrases du document. La particularité de cette méthode est de prendre en compte les liens sur la diagonale pour segmenter.

Le paramétrage de notre méthode influe sur les résultats obtenus, cependant la fourchette de valeurs possibles donnant les meilleurs résultats est assez large. En effet, même si prendre une zone matricielle égale à la moyenne de la taille des segments donne les meilleurs résultats, le fait de prendre une taille jusqu'à quatre fois plus grande ne fait pas chuter les résultats. Le seuil de similarité a les mêmes caractéristiques même si la valeur idoine est 0,5 ; une fourchette est tolérée.

Le TextTiling donne des résultats passables sur le corpus scientifique, dus au fait que la comparaison se fait par blocs. La méthode basée sur la cohérence pallie à ce manque car elle

prend en compte les phrases et n'a donc pas besoin de blocs. Combinée à la méthode de Rocchio, la méthode basée sur la cohérence est améliorée. Grâce aux scores de Rocchio un filtrage des minimums locaux est mis en place et permet d'éliminer ceux ne correspondant pas à un changement de thème.

Les paramètres de notre méthode bien qu'agissant sur les résultats ne sont pas restrictifs et permettent une adaptabilité quant à la diversité des textes.

7 Evaluation Finale

7.1 Le corpus juridique

La méthode choisie pour faire le découpage thématique du corpus juridique est celle de l'adaptation du TextTiling. Les résultats obtenus sont donnés dans le tableau 6.

		F-score
lois	simple	0,248967
	fenêtre 1	0,249402
	fenêtre 2	0,374646

TAB. 6 – Résultats obtenus sur le corpus juridique avec l'adaptation du TextTiling

7.2 Le corpus politique et scientifique

Sur le corpus scientifique et sur le corpus politique nous avons soumis 3 méthodes. Dans la première soumission (1) nous avons utilisé seulement l'apprentissage basé sur la formule de Rocchio avec un seuil de 0.15 pour le corpus politique et de 0.125 pour le corpus politique. La seconde soumission (2) est basée sur la méthode de cohérence avec une taille de fenêtre de 5 avec un seuil de 0,3 pour le corpus scientifique et d'une fenêtre de 15 avec un seuil de 0,5 pour le corpus politique. Enfin la dernière méthode (3) est une combinaison des deux précédentes où les informations d'apprentissages sont prises en compte lors de la sélection des ruptures dans la méthode de cohérence.

		1	2	3	Moyenne DEFT06
politique	Simple	0,2743	0,1345	0,1802	0,1814
	fenêtre 1	0,3869	0,3078	0,2883	0,3030
	fenêtre 2	0,4606	0,4021	0,3514	0,3945
scientifique	Simple	**0,1658**	0,1607	0,1590	0,1150
	fenêtre 1	0,2996	0,2789	**0,3079**	0,2196
	fenêtre 2	0,3557	0,3466	**0,3809**	0,2873

TAB. 7 – Résultats (Fscore) obtenus sur le corpus politique et sur le corpus scientifique

Sur le corpus politique, la méthode par apprentissage fournit les meilleurs résultats, en effet c'est sur corpus que les phrases de ruptures ont des vocabulaires spécifiques. Sur le corpus scientifique la méthode par apprentissage fournit le meilleur Fscore simple, la combinaison fournit les meilleurs résultats pour les autres Fscore. Nous remarquons que la méthode introduite dans cet article donne majoritairement des résultats supérieurs à la moyenne de

DEFT06 or cette méthode est non supervisée et peut donc être appliquée avec peu de connaissance à priori sur le corpus. Les résultats montrent, de plus, qu'il est possible d'ajouter des informations provenant d'un apprentissage pour améliorer les résultats de cette méthode.

8 Conclusion et perspectives

Pour segmenter chacun des trois corpus de la tâche proposée par DEFT'06, nous avons utilisé plusieurs méthodes : apprentissage, adaptation du TextTiling et une méthode basée sur la notion de cohérence. Cette dernière méthode est basée sur une zone de comparaison intermédiaire (on ne fait pas des comparaisons de similarités entre zones adjacentes ou en tenant compte de toutes les zones). Nous avons proposé un nouveau calcul de similarité basé sur les liens et la cohérence. Quant à l'extraction des segments thématiques, elle se fait en fonction des valeurs de cohérence données par les phrases.

Certaines améliorations et perspectives sont envisageables. Dans un premier temps, il serait intéressant d'évaluer la généricité de la méthode que nous venons de proposer. Celle-ci pourrait également être améliorée grâce au calcul matriciel et par un lissage.

Cette méthode pourrait être appliquée à d'autres contextes, comme celui de la segmentation vidéo (Haïdar 2005). De même, pourquoi ne pas combiner la méthode avec des approches linguistiques (Charolles 1997) ? Enfin, il serait intéressant d'utiliser cette méthode sur le corpus juridique en tenant compte de ses spécificités, à savoir utiliser des blocs de textes compris entre deux apparitions consécutives de « Article X ». Une autre alternative est d'évaluer le nombre de liens de chaque phrase « Article X ». Une phrase ayant beaucoup de liens nous indique une continuité de thèmes alors qu'une phrase ayant peu de liens nous indique une rupture.

Références

Amini M., Zaragoza H. & Gallinari P. (2000). Learning for sequence extraction tasks. In *Proceedings*

Brants T., Chen F., Tsochantaridis I. (2002). Topic-based document segmentation with probabilistic latent semantic anaysis. In *Proceedings of CIKM'02*, McLean, Virginia, USA.

Brouard C. (2002) RELIEFS : un système d'inspiration cognitive pour le filtrage adaptatif de documents textuels, *Revue des Sciences et Technologies de l'Information*, vol7, no1/2, 157-182.

Caillet M., Pessiot J.-F., Amini M., Gallinari P. (2004). Unsupervised learning with term clustering for thematic segmentation of texts. In *Proceedings RIAO'04*, Avignon, France.

Callan (1994) Passage-level evidence in document retrieval. Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Charolles, M. (1997). L'encadrement du discours - Univers, champs, domaines et espace. *Cahier de recherche linguistique, 6*.

Choi F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the1st Meeting of the North American Chapter of the Association for Computational Linguistics*, USA.

Chuang S.-L. & Chien L.-F. (2004). A practical web-based approach to generating topic hierarchy for text segments. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management table of contents*, p. 127–136, Washington, D.C, USA.

Haidar S., Joly P., Chebaro B. (2005) Style Similarity Measure for Video Documents Comparison, *4th Int. Conf. on Image and Video Retrieval (CIVR2005), Singapore, 20/07/2005-22/07/2005, Springer-Verlag GmbH, LNCS Vol. 3568, ISBN: 3-540-27858-3, ISSN: 0302-9743, p. 307-317, juillet 2005.*

Hearst M.A. (1997) TextTiling: Segmenting Text into Multi-paragraphe Subtopic Passage, Actes de *Computational Linguistics*, 33-64.

Reynar. J.C. (1998) *Topic segmentation: Algorithms and applications.* Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

Kan M.Y., Klavans J.L., McKeown K.R.. (1998). Linear segmentation and segment significance. In *Proceedings of the 6th. Inter~national Workshop of Very Large Corpora (WVLC-6),* pages 197-205, Montreal, Quebec,Canada, August.

Marti A. Hearst M.A. (1994) Multi-paragraph segmentation of expository text. In *Proceedings of the ACL' 94.* Las Crees, NM.

Mekhaldi D., Lalanne D., Ingold R. (2004) Using bi-modal alignment and clustering techniques for documents and speech thematic segmentations. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management table of contents*, p. 69–77, Washington, D.C, USA.

Salton, A., Buckley C. (1993) Approaches to passage retrieval in full text information systems, *ACM SIGIR Conference on Research and Development in Information Retrieval*

Wilkinson (1994), Effective retrieval of structured models, ACM SIGIR conference on Research and development in information retrieval, 311-317.

Zobel, Moffat, Wilkinson, Sacks-Davis (1994), Efficient retrieval of partial documents, *Information Processing and Management,* vol. 31, n°3, 361-377.

RIAO'2000, Paris, France.

Summary

We introduce in this paper methods we used for the DEFT'06 evaluation campaign in order to automatically recognize thematic segments of French documents on different domains. In this campaign, we evaluate classic methods used in thematic segmentation such as TextTiling or learning and we also introduce a new method based on the notion of coherency.