HAL
open science

Segmentation thématique par calcul de distance thématique

Alexandre Labadié, Jacques Chauché

To cite this version:

Alexandre Labadié, Jacques Chauché. Segmentation thématique par calcul de distance thématique. EGC: Extraction et Gestion des Connaissances, Jan 2007, Namur, Belgique. pp.355-366. lirmm-00161992

HAL Id: lirmm-00161992

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00161992

Submitted on 12 Jul 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Segmentation thématique par calcul de distance thématique

Alexandre Labadié*, Jacques Chauché*

* LIRMM, Université Montpellier 2
UMR 5506
161 rue Ada
34392 Montpellier Cedex 5 - France
alexandre.labadie@lirmm.fr,
jacques.chauche@lirmm.fr

Résumé. Dans cet article, nous présentons une approche de la segmentation thématique fondée sur une représentation en vecteurs sémantiques des phrases et des calculs de distance entre ces vecteurs. Les vecteurs sémantiques sont générés par le système SYGFRAN, un analyseur morpho-syntaxique et conceptuel de la langue française. La segmentation thématique s'effectue elle en recherchant des zones de transition au sein du texte grâce aux vecteurs sémantiques. L'évaluation de cette méthode s'est faite sur les données du défi DEFT'06.

1 Introduction

Le volume toujours plus important de textes rend l'exploitation de ces derniers par des méthodes automatiques de plus en plus complexes. Face à ce problème, la segmentation thématique offre la possibilité d'isoler dans un texte, des segments cohérents du point de vue de leur contenu informationnel. Ainsi, d'autres tâches telles que le résumé automatique ou la recherche d'information par exemple s'en trouve simplifiées. Mais l'on peut imaginer des tâches plus spécifiques telles que la création automatique de table des matières ou de plans à partir d'un gros volume de données non structurées. Nous présentons ici une approche originale de la segmentation thématique en nous appuyant sur les données du défi DEFT'06, Azé et al. (2006).

Pour son édition 2006, DEFT a fixé comme tâche de retrouver les différents segments thématiques d'un grand volume de textes. Trois catégories de textes nous ont été soumises :

- un ensemble de discours politiques.
- un ensemble d'articles de loi.
- un extrait d'un livre à teneur scientifique.

Chacune de ces catégories a été divisées en deux corpus distincts :

- Un corpus d'apprentissage, fourni au début du défi avec les segments thématiques étiquetés, afin d'entraîner nos méthodes.
- Un corpus de test, fourni à la fin du défi, sur lequel nous avons été évalués.

Un calcul de $Fscore$ sur les phrases frontières rapportées par les méthodes a permis l'évaluation des résultats. Les modalités du calcul du $Fscore$, et du couple rappel / précision qui lui est lié, dans le cadre de ce défi sont explicités par Azé et al. (2006).

La tâche de segmentation thématique peut être assimilée à la détection de frontières. Retrouver les segments thématiques au sein d'un texte, revient à retrouver la première phrase (ou la dernière) de chacun de ces segments. Cette phrase jouerait ce rôle de frontière, si toutefois l'épaisseur de la frontière se limite à la phrase (hypothèse fondamentale de l'évaluation).
Les trois catégories de texte sont grandement différentes, et posent donc des problèmes différents. Dans le cas du corpus de la catégorie scientifique (que nous appellerons corpus « scientifique » par la suite) la segmentation thématique consiste à retrouver les différents paragraphes / chapitres du livre. Il nous faut regrouper les articles appartenant au même texte de loi dans le cas du corpus de la catégorie juridique (que nous appellerons corpus « juridique » par la suite). Enfin le corpus de la catégorie discours politiques (que nous appellerons corpus « discours » par la suite) pose lui un double problème :

- Il faut séparer entre eux les différents discours du corpus.
- Au sein même des discours, il faut retrouver les frontières entre les thèmes abordés par l'orateur.

Nous sommes donc devant une triple tâche (voire quadruple si l'on considère la double tâche imposée par le corpus « discours »).
Dans cet article, nous avons toutefois cherché à aborder cet ensemble de tâches complexes sous un angle unique et original, celui de la cohésion sémantique au sein d'un même thème, cohésion que nous chercherons à caractériser.
Après avoir brièvement décrit quelques-unes des méthodes non supervisées les plus courantes à l'heure actuelle dans le domaine de la segmentation thématique, nous présenterons les différentes étapes de notre démarche, depuis la génération des vecteurs sémantiques jusqu'à l'identification des phrases frontières. Nous finirons sur une analyse de nos résultats dans le cadre de la campagne DEFT'06.

2 Méthodes de segmentation thématique non supervisées

Les méthodes de segmentation thématique non supervisées qui ne nécessitent donc ni apprentissage, ni règles, se basent principalement sur la notion de cohésion lexicale observée au travers de la répétition de termes. Par terme, on entend l'unité lexical minimum porteuse de sens, à savoir un mot la plupart du temps, mais parfois un groupe de mots (une collocation), tel que « cul de sac » par exemple.
On peut regrouper ces méthodes en trois grandes familles que nous allons présenter ici.

2.1 Segmentation à partir de mesure de similarité entre segments de texte

Les méthodes de segmentation à base de similarité considèrent les différentes portions de texte du document à traiter comme autant de vecteurs. Les composantes des vecteurs étant, dans la plupart des cas, les fréquences d'apparition des termes au sein de la portion de texte, après que l'on ait retiré les termes inutiles (termes jugés comme peu porteurs de sens) de celle-ci. Parfois, cette fréquence des termes est pondérée par un IDF (Inverse Document Frequency), pour renforcer l'importance des termes supposés thématiquement saillants.
L'objectif de ces méthodes est donc de mesurer la proximité ou l'éloignement des portions de texte étudiées grâce à l'angle que forment leurs vecteurs représentatifs. Elles s'appuient donc

en général sur le cosinus de cet angle, qu'elles considèrent comme un indice de similarité. La similarité est ensuite exploitée de diverses manières. Choi (2000), par exemple, utilise la similarité pour effectuer un classement local et cette approche a retenu notre attention.
Ces méthodes, notamment l'algorithme c99 de Choi (2000), sont, à l'heure actuelle, parmi les plus performantes.
Ces méthodes bien qu'efficaces deviennent rapidement inutilisables à mesure que le volume de données augmente. En effet, ces méthodes s'appuient sur des matrices de similarité entre phrases. Cette approche est donc difficile à mettre en ¡uvre dans le cas de masses de données volumineuses telles que celles issues du défi DEFT'06 (pour un volume de 400000 phrases on obtient : $400000 * 400000 = 1.6 * 10^{11}$ entrées dans la matrice ; même en utilisant la symétrie de la matrice pour diviser par deux le nombre d'entrées, ce dernier reste trop élevé).

2.2 Segmentation à partir de représentation graphique de répétition de termes

En passant par une représentation graphique des termes, il est plus facile de visualiser leur répartition le long du document étudié. Ainsi, la méthode du nuage de points, présentée par Helfman (1994), emploie cette représentation pour la recherche d'informations. Le principe est de positionner sur un graphique chaque occurrence des termes du document (les termes vides de sens ayant bien entendu été retirés au préalable). Ainsi, un terme apparaissant à une position i et une position j du texte, sera représenté par les 4 couples (i, i), (i, j), (j, i) et (j, j). Les portions du document où les répétitions de termes sont nombreuses apparaîtront alors sur le graphique comme les zones de forte concentration de points.
Cette approche visuelle de la représentation d'un texte a été reprise et adaptée à la segmentation thématique par Reynar (1998) dans son algorithme DotPlotting. L'idée est d'identifier les segments thématiquement cohérents sur le graphique en cherchant les limites des zones les plus denses. La densité d'une région du graphique est calculée en divisant le nombre de points présents dans la région par l'aire de cette dernière. L'objectif de DotPlotting est d'isoler les segments thématiques soit en maximisant leur densité, soit en minimisant la taille des zones « vides » entre les segments. On notera que, dans son principe, cette méthode est très proche de l'algorithme c99 de Choi (2000) et donne des résultats proches même si elle est un peu moins efficace.
Cette approche a même inspiré des méthodes originales, comme celle proposée par Ji et Zha (2003), qui consiste à remplacer le problème de segmentation thématique par un problème de segmentation d'image. Cette méthode utilise une technique de diffusion anisotropique (détection des contours par lissage d'une image) sur la représentation graphique de la matrice de distance afin de renforcer les contrastes entre les zones denses et les frontières.
Comme les précédentes approches, ces méthodes montrent vite leur limite face à de gros volume de texte.

2.3 Segmentation à partir de chaînes lexicales

La segmentation à base de chaînes lexicales relie les occurrences multiples des termes dans un document et estime qu'une chaîne est rompue si la distance entre deux occurrences du même terme est trop importante. Cette distance est généralement exprimée en nombre de

phrases.
Ainsi, la méthode $Segmenter$ présentée par Kan et al. (1998), procède selon ce principe pour effectuer une segmentation thématique du document étudié. On notera tout de même une subtilité. La distance à partir de laquelle l'algorithme considère qu'il y a rupture dépend de la catégorie syntaxique du terme impliqué dans le lien.
Une autre approche fondée sur les chaînes lexicales est proposée par Hearst (1997) avec son algorithme $Text\ Tilling$. Un score de cohésion est attribué à chacun des blocs de texte en fonction du bloc qui le suit. Il est quant à lui calculé sur la base d'un premier score dit « lexical » attribué à chaque paire de phrases en fonction de la paire de phrases qui la suit. Ce score lexical est lui même calculé à partir des paramètres que sont le nombre de termes en commun, de termes nouveaux et de chaînes lexicales actives dans les phrases considérées. Le score de chaque segment de texte est alors le produit scalaire normalisé des scores de chacune des paires de phrases qu'il contient. Si un segment présente un score très différent des segments précédents et suivants, alors la rupture thématique se situe au sein de ce segment.
Ces méthodes ne résolvent pas le problème de la taille variable des frontières et / ou de la localisation précise de ces dernières.

Outre les remarques exprimées sur les limites de ces méthodes par rapport à la tâche demandée, toutes ces approches ont ceci en commun qu'elles s'appuient sur la cohésion lexicale[1] supposée des segments thématiques. Or il est tout à fait possible que deux portions d'un texte aient peu de termes en commun (et donc une faible cohésion lexicale) tout en véhiculant le même contenu informationnel. Même s'il y a eu des tentatives de trouver une cohésion sémantique au sein des textes, grâce notamment à l'adjonction de la LSA à certaines méthodes suscitées (Choi et al. (2001)), la base de ces approches reste très « sac de mot ».
Or un texte est composé d'unités syntaxiques, qui sont également sémantiques, et dont la granularité est supérieure au mot : les phrases. Nous explorons donc ici une méthode pouvant tenir compte de la sémantique d'une phrase.

3 Segmentation thématique : utilisation de vecteurs sémantiques pour la détection de frontière

Si l'on considère que le thème d'un texte est « ce dont parle le texte » et que le sens du texte est le contenu conceptuel de ce dernier, alors il est aisé pour l'humain d'établir un lien entre rupture thématique et rupture sémantique.
Notre approche s'attache à mettre en évidence ce lien supposé entre la structure thématique et la structure sémantique d'un texte. Pour ce faire nous nous sommes appuyés sur une représentation vectorielle des phrases du texte ainsi que sur un certain nombre d'hypothèses sur la manière dont un texte est organisé, notamment en français.

3.1 Prétraitement du texte, SYGFRAN et vecteurs sémantiques

Pour travailler sur les différents corpus de DEFT'06, nous nous sommes appuyés sur une représentation vectorielle des phrases qui nous est fournie par l'analyseur morphosyntaxique

[1] tel que la décrivent Morris et Hirst (1991)

SYGFRAN (Chauché 1984). Le principe est de projeter un terme ou un groupe de termes dans un espace de concepts de dimension finie. Les concepts de cet espace sont issus d'un thésaurus « à la Roget ». Dans notre cas, il s'agit du thésaurus Larousse (1992) qui comprend 873 concepts organisés sur 4 niveaux.
L'analyseur SYGFRAN construit, pour une phrase donnée, un arbre syntaxique. Les feuilles de cet arbre sont les différents termes de la phrase, les n¡uds de l'arbre sont le regroupement des termes de la phrase en groupe ou proposition. La racine de l'arbre représente donc la phrase elle-même. Chaque n¡ud de l'arbre se voit attribuer un vecteur sémantique qui est une combinaison linéaire des vecteurs sémantiques de ses fils. Ainsi prenons par exemple la phrase « Le calcul du sens, qui dépend de la structure syntaxique, utilise une forme vectorielle. » (figure 1). Dans cette phrase le terme « calcul » peut ramener à plusieurs concepts : les mathématiques, la médecine (calcul biliaire), le comportement (quelqu'un de calculateur), entre autres. Toutefois la présence du terme « vectorielle » dans le groupe verbal va confirmer que nous parlons bien de mathématiques ici. La structure syntaxique entre aussi en ligne de compte dans le calcul du vecteur sémantique, principalement comme élément pondérateur des combinaisons linéaires. Toujours dans la même phrase, le groupe nominal prépositionnel rattaché à « calcul », avec les termes « sens » et « syntaxique » véhicule une forte notion de linguistique. Mais comme ce n'est qu'un groupe nominal prépositionnel son importance est jugée moindre et donc le concept de linguistique sera au final présent dans le vecteur sémantique de la phrase, mais dans une moindre mesure par rapport à celui de mathématiques.

Le calcul du sens qui dépend de la structure syntaxique utilise une forme vectorelle.

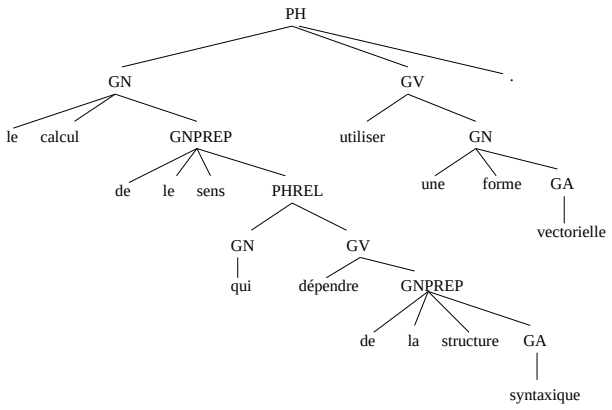

FIG. 1 – *Structure syntaxique*

La représentation vectorielle des phrases, fournie par SYGFRAN, nous permet d'avancer vers une représentation de groupes de phrases, de segments, dans leur intégralité. Il nous faut toutefois nous doter de quelques outils supplémentaires.

3.2 Postulat sur l'organisation thématique d'un texte

En langue française, comme dans toutes les langues, la rédaction d'un texte suit un certain nombre de règles, souvent explicites, mais parfois implicites. Nous sommes partis de la constatation selon laquelle lorsqu'une portion de texte quelconque (paragraphe, chapitre, etc.) traite d'un thème particulier, les premières phrases exposent le sujet abordé, lorsque l'on avance dans le texte, on fait de plus en plus face à des exemples ou des illustrations, pour finir par une ou plusieurs [2] phrases de transitions, qui introduisent le thème suivant. Cette structure, relativement classique, est enseignée dès les premières années d'enseignement secondaire et influence donc la rédaction d'une grande majorité de textes, tant elle est « intégrée » dans notre approche de l'écriture.
On peut donc considérer qu'un texte écrit « selon les règles » aura une structure analogue à celle représenté en 2. Ce postulat rejoint les constatations de Chauché et al. (2003).

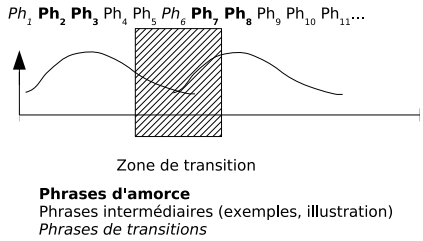

FIG. 2 – *Structure thématique d'un texte*

3.3 Centroïde d'un segment

En partant du postulat, précédent nous avons décidé de représenter un segment thématique non pas par l'ensemble des vecteurs sémantiques (et donc des phrases) qui le composent, mais par un centroïde dont le calcul accordera plus d'importance aux premières phrases qu'aux dernières. Le vecteur centroïde est un barycentre dont les composantes sont calculées selon la méthode de Leibniz. Les dimensions de l'espace étant connues (les vecteurs sémantiques comprennent 873 composantes), nous avons pour $j = 1$ à $j = 873$, n nombre de vecteurs composant le segment thématique, A l'ensemble de ces vecteurs (A_i étant le $i$ème vecteur du segment dans l'ordre d'apparition et x_{j,A_i} la $j$ème composante du vecteur A_i) :

$$x_{j,C} = \frac{\sum_{i=1}^{n} a_i x_{j,A_i}}{\sum_{i=1}^{n} a_i} \tag{1}$$

avec C le vecteur centroïde du segment thématique, $x_{j,C}$ la $j$ème composante du vecteur C et $a_i = n + 1 - i$. Ainsi la pondération a_i qui détermine l'importance que l'on accorde au vecteur courant dans le calcul du barycentre sera égale à n pour le premier vecteur et à 1 pour le dernier, ce qui va dans le sens du postulat que nous avons énoncé plus haut.

[2] mais généralement un petit nombre

3.4 La distance thématique

Afin de pouvoir mesurer la différence thématique entre deux phrases, deux centroïdes ou encore entre une phrase et un centroïde il nous faut disposer d'une fonction similarité ou d'une distance. Nous avons choisi d'adopter la distance thématique présentée par Lafourcade et Prince (2001). Ainsi, si X et Y sont deux vecteurs, D_A étant la distance thématique recherchée, on a :

$$D_A = \arccos(\cos(\widehat{X,Y})) \tag{2}$$

La distance D_A est donc la distance angulaire entre les deux vecteurs X et Y exprimée en radians. Classiquement, le cosinus fait office de mesure de similarité en TALN. Le choix d'appliquer l'arc cosinus pour retrouver la distance angulaire s'est fait pour deux raisons :

- Elle correspond à une distance et présente donc l'avantage d'être réflexive, symétrique et de respecter l'inégalité triangulaire.
- L'$arccosinus$ est une fonction décroissante par rapport à la similarité, mais surtout fortement non linéaire pour des valeurs d'angles faibles (inférieur à $\frac{\pi}{4}$), alors qu'elle se comporte de manière quasi linéaire pour les valeurs d'angles élevées. Ainsi on obtient une plus grande finesse d'analyse lorsque deux phrases sont sémantiquement proches.

En nous appuyant sur ces différents outils, nous pouvons maintenant nous attacher à détecter les zones de transitions au sein d'un texte.

3.5 Détection des zones de transition

Afin de détecter les zones de transition abordées plus haut, nous faisons glisser une fenêtre le long du texte et attribuons à la phrase centrale de la fenêtre une valeur qui correspond à la distance thématique entre le centroïde calculé à partir des phrases précédant la phrase centrale dans la fenêtre (la phrase centrale exclue), et le centroïde calculé à partir de toutes les phrases suivant la phrase centrale (cette dernière étant cette fois incluse dans le centroïde comme le montre la figure 3).

Nous nous sommes servi du corpus d'apprentissage fourni par DEFT'06 plus comme un corpus de calibrage que d'apprentissage. Ainsi la taille de la fenêtre est de deux fois la taille moyenne d'un segment calculée sur le corpus d'apprentissage, une taille est calculée par type de texte. On suppose donc que le corpus de test, sensé être jumeau du corpus d'apprentissage, présentera les même caractéristique.
Une fois la distance thématique estimée, on la compare avec un seuil à partir duquel on considère qu'il y a de fortes chances pour que la phrase fasse partie d'une zone de transition.
Ce seuil est calculé à partir de chacun des corpus d'apprentissage comme suit :

1. Sur chaque corpus nous avons calculé les distances qui séparent chaque segments successif.
2. Nous avons calculé la moyenne et l'écart type de ces distances.
3. Le seuil est égal à la moyenne moins une fois l'écart type.

L'usage de la plus petite distance observée sur le corpus comme valeur seuil a été envisagé, mais sur un tel volume de données traité il y a forcement des accidents et des valeurs particulières. Utiliser la moyenne n'aurait pas forcément été judicieux (éliminant trop de solutions et

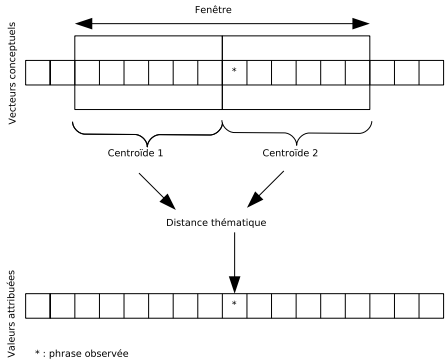

FIG. 3 – *Attribution d'une valeur de distance thématique à chaque phrase*

faisant ainsi chuter le rappel). En utilisant un seuil égal à la moyenne moins l'écart type, on se prémunit des valeurs aberrantes qui pourraient survenir tout en étant moins restrictif que si l'on utilisait la moyenne seulement. Bien entendu, cela implique que nous supposons que le phénomène suit une loi normale. On notera que sur chacun des corpus ce seuil est proche de $\frac{\pi}{4}$, malgré la différence de structure et de discours qu'il existe entre les corpus.
Au final, on obtient deux tableaux, l'un contenant des distances thématiques, l'autre des valeurs booléennes indiquant pour chaque phrase si elle fait partie d'une zone de transition ou non. Toujours dans un souci d'éviter les valeurs singulières, on élimine d'office toutes les phrases marquées comme zone de transition potentielle qui seraient isolées.
Il nous reste à déterminer au sein de cette zone de transition quelles sont les phrases qui constituent vraiment une amorce de segment thématique. Pour ce faire, nous procédons de manière différente selon les corpus.

3.6 Les corpus scientifique et discours et la notion de phrase charnière

Toujours en nous appuyant sur la conception « classique » de la rédaction en langue française, nous avons émis l'hypothèse que pour qu'un texte soit bien construit, il doit comporter des phrases de transition ou phrases charnières (à ne pas confondre avec la zone de transition qui englobe la phrase de transition et les phrases adjacentes) entre chaque portion de texte thématiquement cohérente. Elles ont la particularité d'être la plupart du temps peu porteuses de thème, servant avant tout de lien logique entre deux parties d'un texte. Se trouvant à la frontière entre deux segments thématiques sans avoir de véritable importance thématique, la distance thématique d'une phrase de ce type au centroïde du segment thématique qui la précède doit être proche de la distance au centroïde du segment suivant. Nous avons donc attribué à chacune des phrases traitées un score de transition.
Si on désigne par St_i le score de transition de la phrase i alors on a :

$$St_i = 1 - \frac{2 * |D_p - D_s|}{\pi} \tag{3}$$

Où D_p est la distance de la phrase examinée au centroïde du segment thématique précédent et D_s la distance au centroïde du segment thématique suivant. Cette valeur est comprise entre 0 et 1 et se rapproche de 1 à mesure que les distances entre la phrase examinée et les centroïdes des segments thématiques adjacents se rapprochent. Si les deux distances sont égales (et donc que la phrase centrale est équidistante des deux segments thématiques) elle vaut 1. Si, au contraire, une des distances vaut $\frac{\pi}{2}$ et l'autre 0 (et donc que la phrase centrale est complètement intégrée thématiquement à l'un des segments et pas du tout à l'autre) alors cette valeur vaut 0. Du fait de ces propriétés, nous pouvons utiliser ce score pour pondérer les distances thématiques des phrases suivantes.

A l'étape précédente, nous avons isolé de petites portions de texte susceptibles de contenir la phrase d'amorce d'un segment thématique. Ici, nous allons déterminer quelle phrase au sein de cette portion est la plus susceptible d'être la première phrase d'un segment thématique. Pour ce faire, nous partons du principe qu'une phrase est la première phrase d'un nouveau segment thématique si :

- La distance thématique qui lui a été attribuée est la plus élevée.
- La phrase qui la précède a de fortes chances d'être une phrase de transition.

Comme il est peu probable que ces 2 conditions soient réunies simultanément, nous associons à chaque phrase de la zone de transition, une nouvelle valeur qui est le produit de la distance thématique attribuée à la phrase avec le score de transition de la phrase qui la précède. Il ne nous reste plus qu'à sélectionner le maximum.

3.7 Le corpus juridique et son traitement simplifié

La structure même d'un texte juridique exclut les phrases de transitions entre les articles. Il n'aurait donc pas été pertinent de rechercher ces dernières dans le cadre du traitement du corpus juridique. Toutefois, le corpus « loi » se présente sous la forme d'une succession d'articles tous précédés de la mention « Article X » (le X remplaçant le numéro de l'article afin que ce dernier ne soit pas immédiatement identifiable comme le premier d'une loi).
Nous avons choisi de continuer à chercher les zones de transition selon la méthode présentée plus haut, mais pour déterminer quelle était la phrase d'amorce au sein de ces groupes de phrases nous recherchions simplement la phrase « Article X ».

4 Résultats expérimentaux

Pour évaluer les résultats, l'équipe organisatrice s'est appuyée sur trois calculs de $Fscore$ différents. Si tous donnent la même importance à la précision et au rappel, les trois $Fscores$ se différencient par un certain degré de tolérance à l'erreur. Si le $Fscore$ strict ne se calcul qu'en considérant les phrases ramenées, les $Fscores$ souples de taille 1 et 2 considèrent également les phrases autour de la phrase ramenée (immédiatement adjacentes pour la taille 1 et éloignées d'une phrase pour la taille 2).

Les résultats obtenus sont malheureusement partiels du fait d'un problème de temps lié au pré-traitement du corpus (le corpus « loi » n'étant traité qu'au quart dans la deuxième exécution). La première exécution se base sur une méthode proche de la méthode décrite dans cet

	« D. »	« L. »	« S. »
Rappel strict	0.2	0.17	0.07
Précision stricte	0.06	0.19	0.1
Fscore strict	0.1	0.18	0.08
Rappel souple (taille 1)	0.6	0.17	0.14
Précision souple (taille 1)	0.19	0.19	0.23
Fscore souple (taille 1)	0.29	0.18	0.17
Rappel souple (taille 2)	0.98	0.18	0.14
Précision souple (taille 2)	0.32	0.23	0.24
Fscore souple (taille 2)	0.48	0.21	0.18

TAB. 1 – *Résultats sur les corpus de test*

	« D. »	« L. »	« S. »
Rappel strict	0.32	0.81	0.24
Précision stricte	0.06	0.15	0.05
Fscore strict	0.11	0.26	0.08
Rappel souple (taille 1)	0.79	0.81	0.76
Précision souple (taille 1)	0.16	0.15	0.15
Fscore souple (taille 1)	0.26	0.26	0.25
Rappel souple (taille 2)	0.98	0.95	0.91
Précision souple (taille 2)	0.22	0.19	0.19
Fscore souple (taille 2)	0.36	0.32	0.31

TAB. 2 – *Résultats de la méthode présentée par Chauché (2005) à DEFT'05 appliquée aux corpus de test*

	« D. »	« L. »	« S. »
Rappel strict	0.24	0.25	0.20
Précision stricte	0.19	0.17	0.09
Fscore strict	0.18	.17	0.11
Rappel souple (taille 1)	0.44	0.25	0.39
Précision souple (taille 1)	0.3	0.2	0.17
Fscore souple (taille 1)	0.3	0.2	0.22
Rappel souple (taille 2)	0.55	0.33	0.47
Précision souple (taille 2)	0.38	0.25	0.23
Fscore souple (taille 2)	0.39	0.26	0.29

TAB. 3 – *Moyennes de toute les équipes sur le défi*

article et présentée par Chauché (2005) lors de l'atelier DEFT'05 et se base sur une recherche de maxima locaux plutôt que sur des seuils déterminés à partir d'un corpus d'apprentissage. Toutefois, même partiels, ces résultats nous permettent de faire un certain nombre de constatations :

- Ce qui ressort avant tout de ces résultats, c'est la faible précision de la méthode par rapport à la moyenne des participants. Ce manque de précision peut aisément s'expliquer. En effet, l'objectif de cette méthode est de détecter les zones au sein du texte où le thème change, pas la phrase exacte qui marque ce changement.
- Le recour à un calcul de $Fscore$ souple pour l'évaluation de cette tâche se voit totalement justifié. En effet, si les résultats avec un calcul de $Fscore$ strict sont décevants, dès que l'on prend un tant soit peu de marge, ces derniers accusent une hausse significative. Cette remarque renforce l'idée qu'une frontière thématique est plus une zone floue, qu'une unité bien définie.
- Le très fort rappel obtenu sur le corpus discours laisse supposer que les textes politiques obéissent probablement à un schéma d'organisation thématique. Il doit être possible d'extraire ou d'approximer ce dernier de manière algorithmique.

On notera, qu'à l'exception d'une autre équipe que la nôtre, tous les participants au défi se sont appuyés sur des méthodes numériques dérivées de celle présentées au début de cet articles. Les équipes ayant utilisé des approches sémantiques et structurelles (dont la notre) ce situe dans le milieu de tableau (nous sommes 4ème sur le défi, l'autre équipe étant 5ème). Ces approches, étant encore jeunes comparées à des approches numériques, doivent être perfectionnées et ont donc une grande marge de progression devant elles.

5 Conclusion

Dans cet article, nous avons présenté une méthode originale de segmentation thématique qui s'appuie sur une approche sémantique. Cette dernière a déjà été testée sur différents domaines. D'abord en catégorisation de texte, où elle a donné de bons résultats, puis lors de la précédente édition de DEFT, pour identifier des auteurs, où elle a été moins performante. Autour de cette représentation plus sémantique du texte, nous avons étudié une méthode intégrant des contraintes stylistiques pour segmenter thématiquement le texte.

Nous regrettons de n'avoir pu être évalués sur un jeu de données complet, toutefois les résultats obtenus, même partiels, nous laissent entrevoir des possibilités que nous allons explorer en dehors du cadre parfois restrictif d'une situation d'évaluation. Notre classement en milieu de tableau lors de l'atelier DEFT'06 (4ème sur 7) prouve que si notre approche n'est pas aussi efficace que les approches à base de méthodes numériques, elle reste viable, et mérite d'être approfondie.

Même si l'utilisation d'un $Fscore$ souple permet d'avoir une meilleure vision de l'efficacité des méthodes, le découpage même des textes peut être sujet à contestation. La notion de thème telle qu'elle est abordée dans le défi DEFT'06, à savoir l'idée directrice d'un segment de texte, est très subjective. Peut-on affirmer que les différents paragraphes du corpus scientifique forment bien des segments thématiques distincts ? Le découpage des discours politiques est-il approprié ? Sans mettre en doute la compétence des experts qui ont préparé ce corpus, d'autres experts auraient-ils découpé les corpus de la même manière ?

Il pourrait être instructif de procéder à l'évaluation autrement, en proposant par exemple à des

experts humains d'évaluer les résultats de méthodes automatiques, plutôt que de calibrer ces dernières sur leurs productions (que l'on sait imparfaites et subjectives).

Références

Azé, J., T. Heitz, A. Mela, A. Mezaour, P. Peinl, et M. Roche (2006). Présentation de deft'06 (defi fouille de textes). *Actes de DEFT'06 1*, 3–12.

Chauché, J. *Actes de Coling'84*.

Chauché, J. (1990). Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TA Information 1*, 17–24.

Chauché, J. (2005). Application des vecteurs séemantique à la fouille de texte. *Actes de DEFT'05 1*, 113–124.

Chauché, J., V. Prince, S. Jaillet, et M. Teisseire (2003). Classification automatique de textes à partir de leur analyse syntaxico-sémantique. *Actes de TALN'03*, 55–65.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *Actes de NAACL-00*, 26–33.

Choi, F. Y. Y., P. Wiemer-Hastings, et J. Moore (2001). Latent semantic analysis for text segmentation. *Actes de EMNLP*, 109–117.

Ellman, J. et J. Tait (1999). Roget's thesaurus: An additional knowledge source for textual cbr? *Actes de SGES Internationnal Conference on Knowledge-Based and Applied AI*, 204–217.

Hearst, M. A. (1997). Text-tilling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 59–66.

Helfman, J. (1994). Similarity patterns in language. *Visual Languages*, 173–175.

Ji, X. et H. Zha (2003). Domain-independant segmentation using anisotropic diffusion and dynamic programming. *Actes de ACM/SIGIR Conference of Research and Developpement in Information Retrieval*.

Kan, M., J. L. Klavans, et K. R. McKeown (1998). Linear segmentation and segment significance. *Actes de WVLC-6*, 197–205.

Lafourcade, M. et V. Prince (2001). Synonymie et vecteurs conceptuels. *Actes de TALN'01*, 233–242.

Larousse (1992). *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Paris : Larousse.

Morris, J. et G. Hirst (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics 17*, 20–48.

Reynar, J. C. (1998). *Topic Segmentation: Algorithms and Applications*. Phd thesis, University of Pennsylvania.

Roget, P. (1852). *Thesaurus of English Words and Phrases*. London: Longman.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. *Actes de COLING92*.

Summary

In this article, we present a topic segmentation approach based on a sentence representation by semantic vector and distance calculation between these vectors. The sementic vectors are generated by the SYGFRAN system, a morpho-syntactic and conceptual analyser of the french language. The topic segmentation is done by seeking transition windows in the text using semantic vectors. This approach as been evaluated on the DEFT'06 challenge's datas.