Segmentation thématique par calcul de distance sémantique

Alexandre Labadié*, Jacques Chauché**

* UMR 5506
161 rue Ada
34392 Montpellier Cedex 5 - France
alexandre.labadie@lirmm.fr,
** jacques.chauche@lirmm.fr
http ://www.lirmm.fr/ chauche/

Résumé. Dans cet article, nous présentons une approche de la segmentation thématique basée 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.

1 Introduction

Pour son édition 2006 DEFT nous a donné pour 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 et un extrait d'un livre à teneur scientifique, chacune de ces catégories définissant un des corpus sur lesquels il nous a fallu travailler.
La tâche de segmentation thématique peut être assimilée à de la détection de frontière. Retrouver les segments thématiques au sein d'un texte, c'est 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). Toutefois peut on vraiment considérer dans le cas présent qu'il s'agisse d'une unique tâche ?
En effet les trois catégories de texte (et donc les 3 corpus) sont grandement différentes. 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 « loi » par la suite). Le cas du 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 complexe 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 la phase de prétraitement du texte, à savoir la génération des vecteurs sémantiques. Nous terminerons notre propos en détaillant les méthodes de segmentation thématique à proprement parler.

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. 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 mots au sein de la portion de texte, après que celle-ci a été débarrassée des mots inutiles (mots jugés comme peu porteurs de sens). Parfois, cette fréquence des mots est pondérée par un IDF (Inverse Document Frequency), pour renforcer l'importance des mots 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 la similarité. La similarité est ensuite exploitée de diverses manières. Choi (Freddy Y. Y. Choi 2000), par exemple, utilise la similarité pour effectuer un classement local et cette approche a retenu notre attention.
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 similarités entre phrases, or le volume de données à traiter dans DEFT'06 ne se prête guère à ce genre de représentation ($400000 * 400000 = 1.6 * 10^11$ 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 (Jonathan Helfman 1994) emploie cette représentation pour la recherche d'information. Le principe est de positionner sur un graphique chaque occurrence des termes du document. 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 (Jefrey C. 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 (Freddy Y. Y. Choi 2000).
Cette approche a même inspiré des méthodes originales, comme celle proposée par Ji et Zha (X. Ji and H.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 sur la représentation graphique de la matrice de distance afin de renforcer les contrastes entre les zones denses et les frontières.
Par rapport à la problématique, ces méthodes se situent plus dans une détection des frontières par défaut, plutôt que dans une véritable recherche de ces dernières, qui est la tâche qui nous a été confiée.

2.3 Segmentation à partir de chaînes lexicales

La segmentation à base de chaînes lexicales relie les occurrences multiples des mots dans un document et estime qu'une chaîne est rompue si la distance entre deux occurrences du même mot est trop importante. Cette distance est généralement exprimée en nombre de phrases.
Ainsi, la méthode *Segmenter* présentée par Kan (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 mot impliqué dans le lien.
Une autre approche basée sur les chaînes lexicales est proposée par Hearst (M. A. 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 mots en commun, de mots 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 se basent 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 mots en commun (et donc une faible cohésion lexicale) tout en véhiculant le même contenu informationnel. Même s'il y a eu des tentatives d'intégrer une information de type sémantique, grâce notamment à l'adjonction de la LSA à certaines méthodes suscitées (Choi et al. 2001), la base de l'approche reste très « sac de mot ».
Or un texte est composé d'unités syntaxiques, qui sont également sémantiques, et dont la granularité et supérieure au mot : les phrases. Nous explorons donc ici une méthode pouvant tenir compte de la sémantique d'une phrase.

3 Prétraitement du texte, SYGFRAN et vecteurs sémantiques

La méthode vectorielle que nous présentons ci-dessous peut être utilisée pour représenter aussi bien un mot qu'un ensemble ordonné de mots.

3.1 Vecteur de terme

Définition
Un vecteur sémantique projette un terme donné dans un espace sémantique dont une famille génératrice correspond à un ensemble d'idées.

L'ensemble des idées nécessaires pour former une famille génératrice peut être définie par un thésaurus.

La procédure est la suivante : on projette la totalité des lexies du dictionnaire sur un espace défini à partir d'une famille de concepts « à la Roget ». Pour le Français, les lexicologues du Larousse ont défini une famille de 873 concepts hiérarchisés en 4 niveaux. Sur un plan vectoriel, cela produit un espace à 873 dimensions que l'on admet comme étant de dimension donnée. Les approches « à la Roget » sont relativement nombreuses depuis quelques années, dans la littérature anglo-saxonne. En Français, l'indexation automatique à partir du thésaurus a été proposée à l'origine par nous-mêmes, mais on la retrouve aujourd'hui utilisée dans de nombreux travaux.

Formellement, on considère que tout terme t du dictionnaire est représenté par un vecteur $\vec{t}$ dans l'espace vectoriel considéré, que l'on nommera $\vec{\mathcal{V}}$. On suppose qu'il

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

existe une application qui plonge l'espace lexical linguistique dans l'espace vectoriel engendré par la famille de concepts du thésaurus. Pour des besoins de calcul, seule une version normée $\vec{t}_{nor}$ de ce vecteur est conservée dans l'espace. Comme on ne traite que de vecteurs normés, par convention, on écrira $\vec{t}$ pour désigner le vecteur normé du terme t. Pour cela, on introduit une norme euclidienne sur l'espace vectoriel sémantique.

La majorité des mots,étant polysémique, renvoie à une multiplicité d'idées, ou concepts du thésaurus.

Exemple
Les idées associées au mot *calcul* sont par exemple : Calcul, Opération arithmétique, Maladie et Intention.

L'emploi de ce mot simplement ne permet donc pas de définir sa signification : par exemple, *calcul arithmétique* ou *calcul biliaire*, ou *Il m'a aidé par calcul.*

Cela signifie que le terme doit être représenté, non seulement par la manière dont il est indexé dans le thésaurus, mais aussi par ses différentes significations, qui elles, ont un sens lorsque le mot est utilisé dans une construction (groupe ou phrase).

Le calcul sémantique sur une phrase doit donc incliner le sens du mot "calcul" vers une des significations possibles.

3.2 Vecteur sémantique d'une phrase

Définition
On dira que l'on représente toute *phrase* construite, par un vecteur produit comme une combinaison linéaire de vecteurs sémantiques des *groupes* qui la composent.
On dira que l'on représente tout *groupe* construit, par un vecteur produit comme une combinaison linéaire de vecteurs sémantiques des *termes* qui le composent.

Pour cela on introduit les opérations suivantes :
Somme normée : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ représentant les vecteurs (normés) de deux termes t_1 et t_2.

$$\overrightarrow{(t_1 + t_2)_{nor}} = \frac{\vec{t_1}+\vec{t_2}}{\|\vec{t_1}+\vec{t_2}\|} \tag{1}$$

Remarque : la somme normée n'est pas associative :
$\overrightarrow{(t_1 + t_2 + t_3)_{nor}}$ n'est pas égal à $(\overrightarrow{(t_1 + t_2)_{nor}} + \vec{t_3}))_{nor}$. Par convention, on ne retiendra comme opération de somme que la somme normée, et on omettra dorénavant l'indice 'nor'.

Multiplication par un scalaire : Soit un vecteur $\vec{t}$ normé. Soit λ un scalaire. Le vecteur $\vec{\lambda t}$ est égal à $\lambda * \vec{t}$. Cela signifie que toutes les composantes du vecteur sont multipliées par le scalaire.

Remarque : cette multiplication a pour objectif de renforcer la « présence » du vecteur dans une combinaison linéaire, et ne s'utilise en principe jamais isolément.

Produit terme à terme : Soient deux vecteurs $\vec{t_1}$, et $\vec{t_2}$ normés. Le produit terme à terme des deux vecteurs se définit comme :

$$\overrightarrow{(t_1 * t_2)_{nor}} = \frac{\vec{t_1}*\vec{t_2}}{\|\vec{t_1}*\vec{t_2}\|} \tag{2}$$

où si $a_{p,i}$ est la ième composante de $\vec{t_1}*\vec{t_2}$, et $a_{1,i}$ et $a_{2,i}$ respectivement celles de $\vec{t_1}$, et $\vec{t_2}$, on a :

$$\forall i \in [1, 873], a_{p,i} = a_{1,i} * a_{2,i} \tag{3}$$

»

Par convention, on omettra l'indice *nor* et on appellera par défaut $\overrightarrow{(t_1 * t_2)}$ le produit terme à terme normé.

Distance « angulaire » : La distance selon Salton, servant de mesure de similarité est calculée comme le *cosinus* de l'angle de deux vecteurs.

$$sim(\vec{t_1}, \vec{t_2}) = \cos \widehat{\vec{t_1}, \vec{t_2}} = \frac{\vec{t_1}.\vec{t_2}}{\|\vec{t_1} * \vec{t_2}\|} \tag{4}$$

où « . » est le produit vectoriel classiquement défini. La distance que nous utilisons correspond à une mesure relative à l'angle $\widehat{\vec{t_1}, \vec{t_2}}$. Comme nous ramenons tous les angles considérés à l'espace $[0, \frac{\pi}{2}]$, alors la mesure que nous proposons se calcule par :

$$\delta(\vec{t_1}, \vec{t_2}) = 1 - \cos \widehat{\vec{t_1}, \vec{t_2}} \tag{5}$$

Remarques : Ramener les valeurs de δ à $[0, 1]$ est plus pratique que de mesurer des valeurs entre 0 et $1,67$ radiants. Lorsque deux vecteurs sont totalement divergents (intersection vide), leur angle est de $\frac{\pi}{2}$, et le cosinus vaut 0 : leur distance est maximale et vaut 1. Lorsque ces vecteurs sont très proches, leur angle tend vers 0 , le cosinus tend vers 1 et la distance, vers 0. Tous les vecteurs ont un angle forcément compris entre 0 et $\frac{\pi}{2}$, par construction, et appartiennent au même espace vectoriel.

3.3 Vecteur de groupe

La deuxième propriété du calcul sémantique correspond à une définition différentiée d'un groupe suivant sa structure. Ainsi le sens du groupe "le calcul du sens" est distinct du sens du groupe "le sens du calcul",ces deux groupes ayant rigoureusement les mêmes éléments (le langage naturel n'étant pas commutatif). Comme le mot "sens" est très riche sémantiquement (une vingtaine de sens justement) nous prendrons pour l'exemple de la représentation l'idée associée : Sens.L'id/'e est différente du terme, selon les lexicologues, en ce qu'elle étiquette un champ sémantique. Le terme peut appartenir ou relever de plusieurs champs, en raison de sa polysémie. Dans le sous-espace ayant comme axe *Calcul*, *Intention* et *Sens* les vecteurs associés aux deux groupes précédents seront :

3.4 Calcul du vecteur de phrase

Le calcul d'un vecteur de phrase s'effectue (sur une phrase) en plusieurs étapes à partir de la structure syntaxique :

– La première étape consiste à associer à chaque feuille un vecteur sémantique issu de la lecture d'un dictionnaire (vecteur de terme)

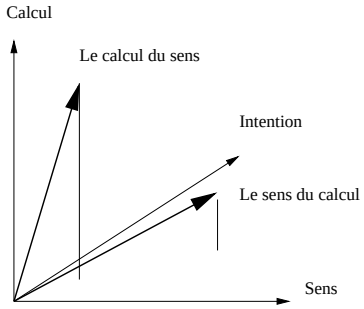

FIG. 1 – Vecteur de groupe

Si un élément à plusieurs sens ou interprétations possibles, le vecteur associé correspond au *centroïde* de l'ensemble des vecteurs associés à chaque interprétation (somme normée de tous les vecteurs indexant ce terme).

– La deuxième étape consiste à calculer récursivement le vecteur associé à chaque groupe.

Le vecteur associé à un groupe est obtenu par une combinaison linéaire des vecteurs associés aux éléments de ce groupe. Les coefficients de cette combinaison linéaire dépendent de la fonction syntaxique de l'élément : gouverneur du groupe, sujet, objet, etc...

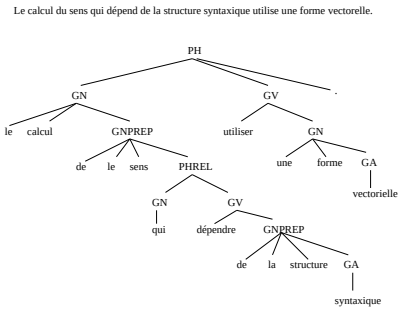

FIG. 2 – Structure syntaxique

– La troisième étape actualise les vecteurs associés aux feuilles. Cette actualisation consiste à effectuer un produit terme à terme du vecteur à actualiser avec le vecteur obtenu du texte.

Cette actualisation terminée un nouveau calcul est effectué. La convergence est très rapide et deux itérations suffisent pour obtenir un vecteur significatif.

4 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 se faire nous avons formulé un certain nombre d'hypothèses sur la manière dont un texte est organisé, notamment en français.

4.1 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, et, alors que 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 (mais généralement un petit nombre) 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 (mais pas la totalité, comme nous le verrons pour le cas du corpus juridique), tant elle est « intégrée »dans notre approche de l'écriture.
On peut donc considérer qu'un texte écrit « selon les règles » aura une structure du type de celui représenté par la figure 3.

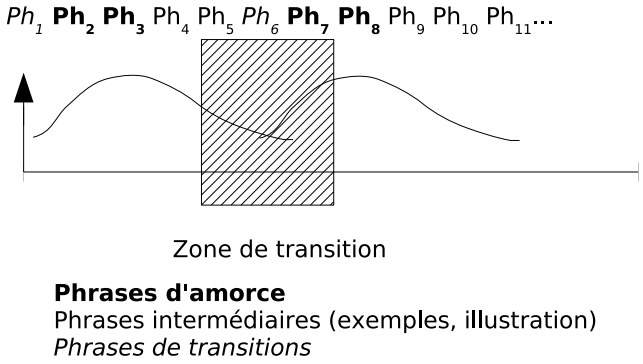

FIG. 3 – Structure thématique d'un texte

Ce postulat rejoint les constatations de Chauché *et al.* (2003).

4.2 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) :

$$x_{j,C} = \frac{\sum_{i=1}^{n} a_i x_{j,A_i}}{\sum_{i=1}^{n} a_i} \tag{6}$$

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 calcule 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.

4.3 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 (Mathieu Lafourcade et Violaine Prince 2001). Ainsi, si X et Y sont deux vecteurs, D_A étant la distance thématique recherchée, on a :

$$D_A = \arccos(\cos(\widehat{X, Y})) \tag{7}$$

La distance D_A étant exprimée ici en radians.

4.4 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 4).

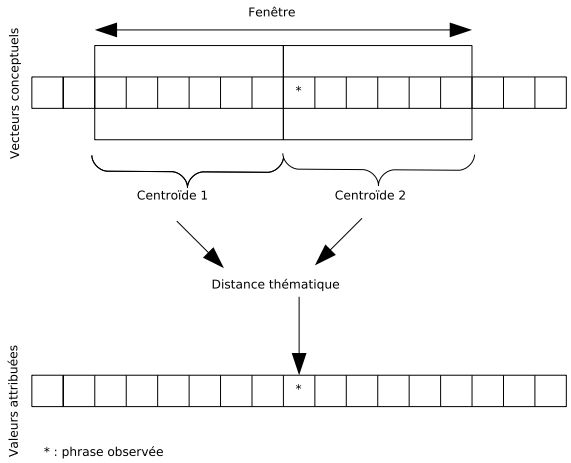

FIG. 4 – Attribution d'une valeur de distance thématique à chaque phrase

La taille de la fenêtre est de deux fois la taille moyenne d'un segment thématique au sein du corpus d'apprentissage, afin de couvrir une majorité de cas.
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 des corpus d'apprentissage. Il en existe un par corpus et il est égal à la moyenne des distances thématiques entre les centroïdes des segments successifs moins 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 juste la moyenne n'aurait pas forcément été judicieux (éliminant trop de solutions et 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 on utilisait la moyenne seulement. On notera que ce seuil a une valeur comprise entre 0.6 et 0.7 radian selon 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.

4.5 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 :

$$St_i = \frac{\frac{\pi}{2} - |D_p - D_s|}{\frac{\pi}{2}} \tag{8}$$

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.

4.6 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 trai-

tement du corpus juridique. Toutefois, nous savions que tous les segments thématiques commençaient par la forme « Article X » (même si tous les « Article X » n'étaient pas des débuts de segments thématiques).
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 ».
Le corpus juridique a bénéficié d'autres traitements spécifiques du fait de son caractère très particulier. Ainsi la fenêtre que nous faisons passer sur le texte lors du calcul des distances thématiques est d'une taille fixe de 20 phrases. En effet si pour un texte « classique » la taille moyenne a un sens, pour les textes juridiques où certains articles remplissent des pages entières alors que d'autres se résument à une phrase, sans qu'il y ait vraiment de norme, la moyenne perd tout son sens. A titre d'exemple sur le corpus juridique d'apprentissage la taille moyenne d'un segment thématique et de 22.45 phrases avec un écart type de presque 13 phrases. La taille de 20 phrases (soit deux segments de 10 phrases) a été choisie de manière empirique, c'est elle qui donnait les meilleurs résultats sur le corpus d'apprentissage.

5 Bilan

	Exécution 1			Exécution 2		
	« D. »	« L. »	« S. »	« D. »	« L. »	« S. »
Rappel strict	0.32	0.81	0.24	0.2	0.17	0.04
Précision stricte	0.06	0.15	0.05	0.06	0.19	0.04
Fscore strict	0.11	0.26	0.08	0.1	0.18	0.04
Rappel souple (taille 1)	0.79	0.81	0.76	0.6	0.17	0.15
Précision souple (taille 1)	0.16	0.15	0.15	0.19	0.19	0.15
Fscore souple (taille 1)	0.26	0.26	0.25	0.29	0.18	0.15
Rappel souple (taille 2)	0.98	0.95	0.91	1	0.18	0.22
Précision souple (taille 2)	0.22	0.19	0.19	0.32	0.23	0.23
Fscore souple (taille 2)	0.36	0.32	0.31	0.48	0.21	0.23

D. : Discours, L. : Loi, S. : Scientifique.

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 pour l'exécution 2 notamment). La première exécution se base sur une méthode proche de la méthode décrite dans cette article et présentée par Jacques Chauché lors de l'atelier DEFT'05 (J. Chauché 2005). 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 des deux méthodes. Ce manque de précision peut aisaiment s'expliquer. En effet l'objectif de ces méthodes est de detecter 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. Car, si les résultats avec un calcul de *Fscore* strict sont décevants, dès que l'on prend un tant soit peu de marge ils grimpent rapidement. Cette remarque renforce l'idée qu'une frontière thèmatique est plus une zone floue, qu'une unité bien définie.
– La méthode utilisée pour la première exécution offre un bon rappel (bien meilleur que celle de la seconde), mais se révèle un peu moins précise (notamment sur le corpus « discours »). On peut en déduire que l'apprentissage d'un seuil, par rapport à la recherche d'un maximum local, s'est révélé beaucoup moins performant sur le corpus « scientifique » que sur le corpus « discours ».

Cette dernière remarque nous laisse supposer que les discours politiques obéissent probablement à un schéma d'organisation thématique (volontaire ou non) qu'il doit être possible d'extraire ou d'approximer de manière algorithmique.

6 Conclusion

Nous souhaitions tester dans le cadre de cette évaluation une approche sémantique qui a déjà été testée dans d'autres 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 deux variantes d'une méthode intégrant des contraintes stylistiques pour segmenter thématiquement le texte.
Nous regrettons de n'avoir pu être évalué 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.
Ainsi on peut regretter la méthode utilisée pour l'évaluation des résultats. Même si l'utilisation d'un fscore souple permet d'avoir une meilleur 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 la tâche, 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

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

J. Chauché (2005), Application des vecteurs sémantique à la fouille de texte. Actes de DEFT'05, pp 113-124.

J. Chauché, V. Prince, S. Jaillet, M. Teisseire (2003) Classification Automatique de Textes à partir de leur Analyse Syntaxico-Sémantique TALN'03 : 10ème Conférence Internationale sur le Traitement Automatique du Langage Naturel, pp. 55-65

F. Y. Y. Choi (2000), Advances in domain independent linear text segmentation, Proceedings of NAACL-00, pp 26-33.

F. Y. Y. Choi, P. Wiemer-Hastings and J. Moore (2001), Latent Semantic Analysis for Text Segmentation, Proceedings of 6th EMNLP, pp 109-117.

J. Ellman, J. Tait (1999) Roget's thesaurus : An additional Knowledge Source for Textual CBR ? Proc. of 19th SGES Int. Conf. on Knowledge-Based and Applied AI. Springer-Verlag. pp 204 - 217.

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

J. Helfman (1994), Similarity Patterns in Language, Visual Languages, pp 173-175.

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

Min-Yen Kan and J. L. Klavans and K. R. McKeown (1998), Linear Segmentation and Segment Significance, Proceedings of WVLC-6, pp 197-205.

M. Lafourcade et V. Prince (2001), Synonymie et vecteurs conceptuels, TALN 2001, pp 233-242.

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

J. Morris et G. Hirst (1991), Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, Computational Linguistics, vol.17, N°1., pp.20-48.

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

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

D. Yarowsky (1992), Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proceeding of 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.