Application des vecteurs sémantiques à la fouille de texte

Jacques Chauché
LIRMM-CNRS et Université Montpellier 2
161 rue Ada, 34395 Montpellier cedex 5
chauche@lirmm.fr

Mots-clefs : analyse syntaxique, analyse sémantique, similitude sémantique

Keywords: syntactic analysis, semantic analysis, sémantic similitude

Résumé L'approche présentée ici se base sur un traitement du contenu syntaxico-sémantique par un analyseur du Français, le système SYGFRAN, pour retrouver un ensemble de phrases appartenant à différents discours du président François Mitterand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Ce traitement se fait par calcul de vecteurs sémantiques de phrases (méthodologie définie dans l'article) et par la définition d'une relation de similitude décrivant l'inclination de vecteurs dont l'inclinaison, ou distance angulaire, est proche. A l'aide de cette relation, des phrases sont attribuées par le système à l'un ou l'autre des auteurs, et l'article indique des F-mesures obtenues sur le premier corpus, dit d'apprentissage, légèrement supérieures à 80%.

Abstract The approach presented here is based on a treatment of the syntactico-semantics contents by an analyzer of French, system SYGFRAN, to find a group of sentences belonging to various speeches of president François Mitterand mixed with a group of sentences belonging to various speeches of president Chirac. This treatment is done by a calculation of semantic vectors of sentences (methodology defined in the article) and by the definition of a relation of similarity describing the inclination of vectors to which the slope, in angular distance, is close. Using this relation, sentences are allotted by the system to one or the other of the authors, and the article indicates the F-measurements obtained on the first corpus (also called training corpus) slightly higher than 80%.

Le défi 2005 organisé pour le congrès annuel TALN consiste à retrouver un ensemble de phrases appartenant à différents discours du président François Mitterand plongées dans un ensemble de phrases appartenant à différents discours du président Jacques Chirac. Les phrases introduites traitent d'une thématique distincte de la thématique retenue pour les phrases des discours de Jacques Chirac. L'approche présentée ici se fonde sur un traitement du contenu par opposition aux traitements habituels basés sur des approches statistiques. Le vocabulaire utilisé par l'un ou l'autre n'aura d'importance qu'à travers les idées qu'il véhicule.

1 Vecteur sémantique

1.1 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" (Roget 1852). Pour le Français, les lexicologues du Larousse ont défini une famille de 873 concepts hiérarchisés en 4 niveaux (Larousse 1992). 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, (Yarowsky, 1992), (Ellman et Tait 1999). En Français, l'indexation automatique à partir du thésaurus a été proposée à l'origine par nous-mêmes, (Chauché 1990), mais on la retrouve aujourd'hui utilisée dans de nombreux travaux (Crestan et al. 2003).

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 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.

1.2 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.

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

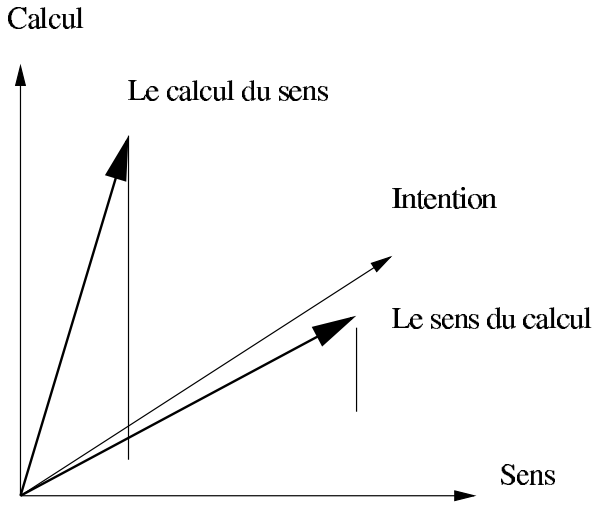
où "." est le produit vectoriel classiquement défini. La distance que nous utilisons correspond à une mesure relative à l’angle $\widehat{\vec{t_1}, \vec{t_2}}$. Comme nous ramenons tous les angles considérés à l’espace $[0, \frac{\pi}{2}]$, alors la mesure que nous proposons se calcule par :

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

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.

1.3 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 :



1.4 Calcul du vecteur de phrase

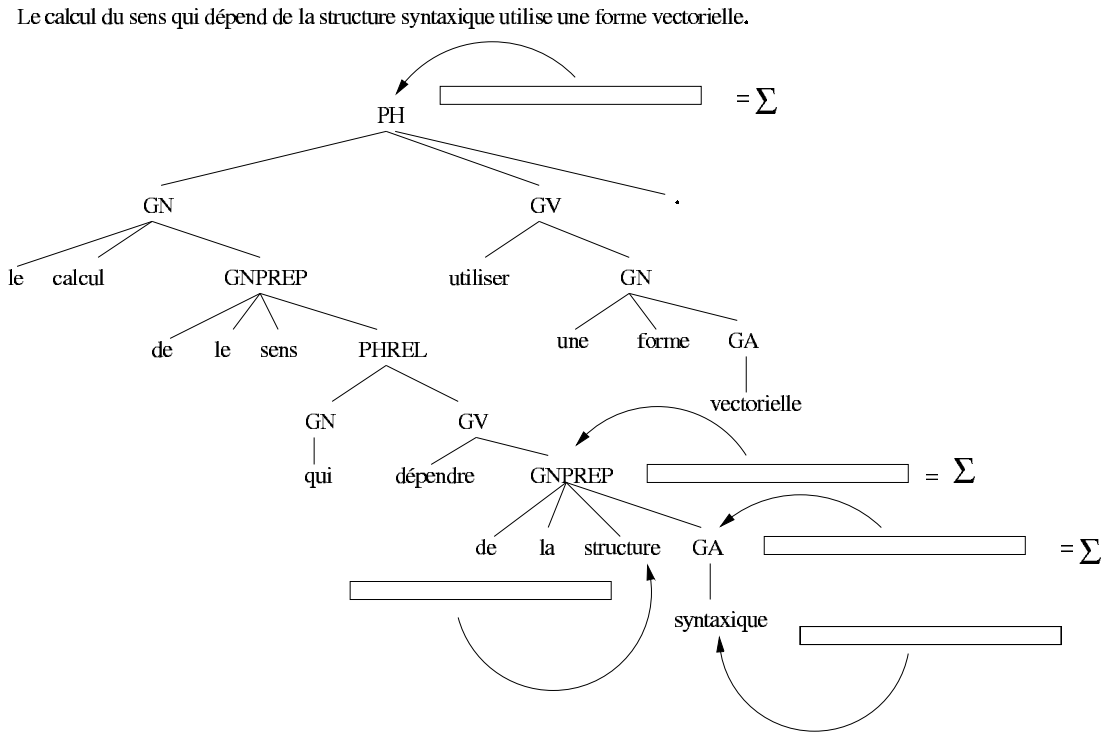
Le calcul d'un vecteur de phrase s'effectue (sur une phrase) en plusieurs étapes à partir de la structure syntaxique :

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

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...



- 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.

1.5 Propriétés du modèle

Le classement s'effectue à partir des vecteurs sémantiques de phrases.

La valeur intrinsèque de la norme d'un vecteur n'est pas significative. Seul compte l'inclinaison de ce vecteur par rapport à une idée ou un autre vecteur donné (la distance angulaire). Aussi tous les calculs se termineront par la normalisation des vecteurs. On ne considérera donc que les points de la shpère unité.

1.5.1 Comparaison d'inclinaison entre vecteurs

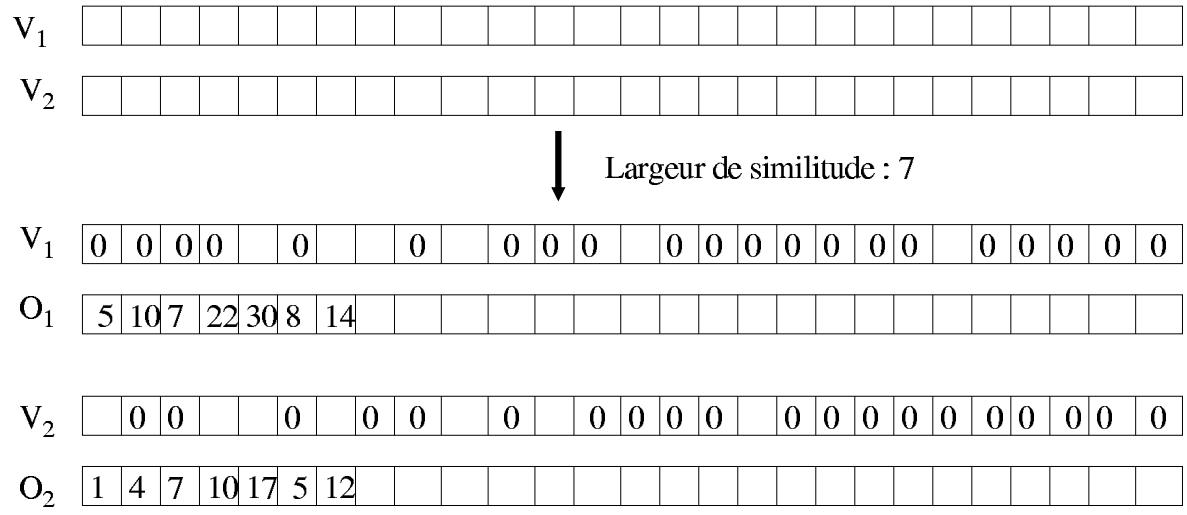
La première mesure de comparaison sera donc la valeur de l'arc séparant deux vecteurs sur cette sphère (Cette mesure sera donc naturellement donnée par la fonction *arcosinus* $(\vec{V_1}, \vec{V_2})$. Comme toutes les composantes de tous les vecteurs sont positives ou nulles nous utiliserons

seulement le produit scalaire. Dans ce contexte un élément sera plus proche d'un autre par rapport à un troisième si le produit scalaire de cet élément avec le troisième a une valeur supérieure au produit scalaire du deuxième avec le troisième.

Le produit scalaire ne rend que très imparfaitement **l'inclination** d'un élément par rapport à l'autre. En effet, l'inclination comprend *l'inclinaison*, mais indique jusqu'à quel point un vecteur "s'assimile" à un autre. Aussi le produit scalaire sera completé par une mesure de *similitude* tenant compte de l'importance relative de chaque idée à l'intérieur de chaque vecteur.

1.5.2 Similitude entre vecteurs d'inclinaison proche, ou mesure d'inclination

Le calcul de la similitude s'effectue sur une largeur donnée. On associe un *vecteur d'indices* à chaque vecteur opérande. Ce vecteur est trié de façon que sa lecture donne un ordre décroissant des composantes du vecteur auquel il est associé. Les composantes du vecteur pour lesquelles l'indice ne se trouve pas dans les premiers éléments du vecteur d'indices sont annulées. Une fois cette opération terminée le nouveau vecteur est renormé. Ensuite la valeur de la similitude correspond à la somme des produits des composantes pondérées par l'écart relatif existant dans les vecteurs d'indices.



$$\alpha_5 = V_1[5] \times V_2[5] \times \frac{1}{1 + \beta \times (1-6) \times (1-6)}$$

$$Sim(V_1, V_2) = \sum \alpha_i$$

Nous avons bien évidemment pour tout vecteur $\vec{V}$ non nul $sim(\vec{V}, \vec{V}) = 1$ et du fait que toutes les composantes sont positives ou nulles :

1.5.3 Propriétés de la similitude

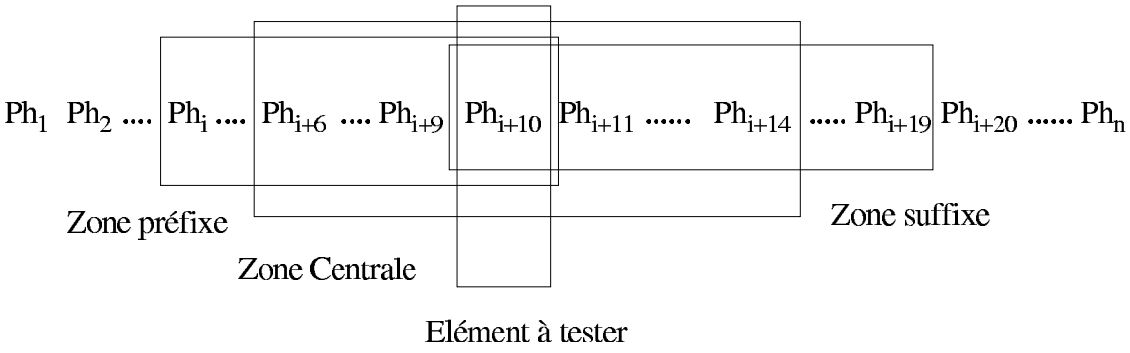
- pour tous vecteurs $\vec{V_1}$ et$\vec{V_2}$ orthogonaux : $sim(\vec{V_1}.\vec{V_2}) = 0$.

- la similitude est également symétrique :

$$sim(\vec{V_1}.\vec{V_2}) = sim(\vec{V_2}.\vec{V_1}) \quad (6)$$

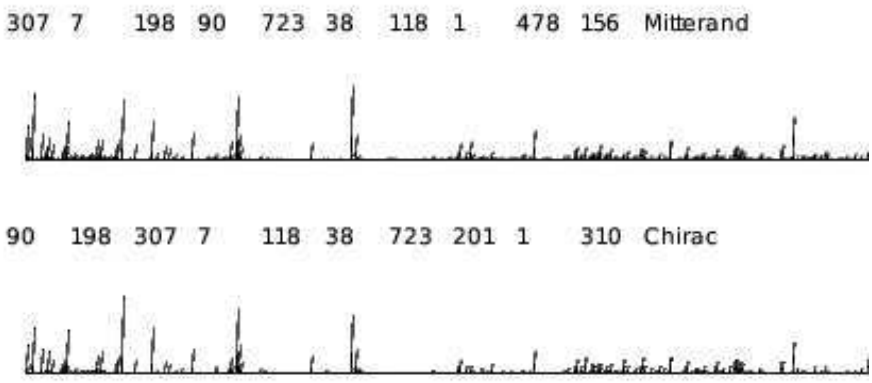
2 Fouille de texte

Le calcul des vecteurs sémantiques s'effectue sur chaque phrase du texte. Le principe de décision pour la sélection d'une phrase est construit sur le calcul moyen des vecteurs associés aux phrase situées à l'intérieur d'une fenêtre. Pour une décision à propos de la phrase Ph_{i+10}les vecteurs concernés seront les centroïdes des trois zones *préfixe,centrale* et *suffixe* telles que définies ci-après.



Le classement des phrases s'effectue par comparaison des différents vecteurs avec des vecteurs spécifiques $\overrightarrow{V_{Chirac}}$et $\overrightarrow{V_{Mitterand}}$, que nous symboliserons par $\vec{V_C}$ et $\vec{V_M}$ respectivement.

> Ces deux vecteurs ont été obtenus en calculant le centroïde des vecteurs des phrases associées à chacun d'eux dans le corpus d'apprentissage. Pour le calcul de ce vecteur, chaque vecteur est affecté d'un poids proportionnel à sa taille (le coefficient utilisé est égal au millième du carré de la longueur en octets).



> Les indices correspondent aux concepts majoritaires dans l'ordre décroisant.

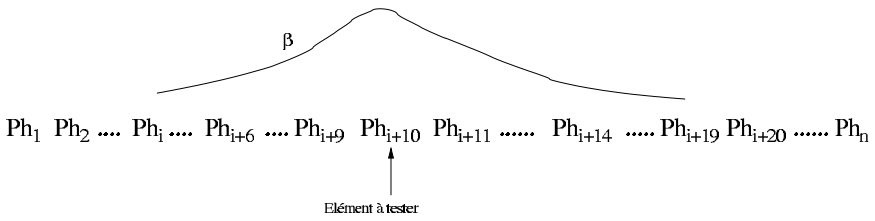
Soient :

- $\vec{V_{ZP}}$ le vecteur de la zone préfixe, que nous représenterons par $\vec{V_{ZP}}$
- $\vec{V_{ZC}}$ le vecteur de la zone centrale, que nous représenterons par $\vec{V_{ZC}}$
- $\vec{V_{ZS}}$ le vecteur de la zone suffixe, que nous représenterons par $\vec{V_{ZS}}$

Le premier filtre compare les produits scalaires $<\vec{V_{ZP}}.\vec{V_C}>$ et $<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ et $<\vec{V_{ZC}}.\vec{V_M}>$ et $<\vec{V_{ZS}}.\vec{V_C}>$ et $<\vec{V_{ZS}}.\vec{V_M}>$.

> Pour qu'une phrase soit candidate au classement comme phrase appartenant au discours de Mitterand il est nécessaire qu'au moins un produit scalaire ($<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_M}>$ ou $<\vec{V_{ZS}}.\vec{V_M}>$) soit supérieur à son correspondant ($<\vec{V_{ZP}}.\vec{V_C}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ ou $<\vec{V_{ZS}}.\vec{V_C}>$).

Dans le cas ou une phrase est candidate au classement un score d'appartenance est évalué. Ce score correspond au nombre de produits scalaires $<V_{Ph_i}.\vec{V_M}>$ supérieur aux produits scalaires $<V_{Ph_i}.\vec{V_C}>$. Dans le calcul du score on fait intervenir un coefficient indiquant la proximité avec la phrase candidate :



Si le score atteint un certain seuil (dépendant de β) et que deux produits scalaires au moins ($<\vec{V_{ZP}}.\vec{V_M}>$, $<\vec{V_{ZC}}.\vec{V_M}>$ ou $<\vec{V_{ZS}}.\vec{V_M}>$) dont le central sont supérieur aux produits scalaires correspondants ($<\vec{V_{ZP}}.\vec{V_C}>$, $<\vec{V_{ZC}}.\vec{V_C}>$ ou $<\vec{V_{ZS}}.\vec{V_C}>$) on effectue un calcul de similitude :

> Le calcul de similitude compare seulement cinq phrases : les deux précédentes, la phrase sélectionnée et les deux suivantes :
>
> pour la phrase sélectionnée :

$$\text{Sim}(\vec{V_M}, V_{Ph_{i+10}}) \text{ et } \text{Sim}(\vec{V_C}, V_{Ph_{i+10}})$$

Si la similitude de la phrase testée par rapport au vecteur représentant le discours de Mitterand est supérieure à la similitude du vecteur testé par rapport au vecteur représentant le discours de Chirac et qu'il en va de même soit pour les deux phrases précédentes soit pour les deux phrases suivantes la phrase testée est attribuée au discours de Mitterand (Nous supposons donc que les parties insérées du discours de Mitterand comportent au moins quatre phrases consécutives).

La recherche des phrases associées au discours de Mitterand se termine par un petit correctif éventuel pour tenir compte de la propriété : *Il n'y a pas de phrase isolée associée au discours de Mitterand.* Ansi si l'on a une configuration comme CCMCC où C désigne une phrase associée au discours de Chirac et M une phrase associée au discours de Mitterand, la phrase associée au discours de Mitterand est désélectionnée. Il en va de même pour l'inverse. Une configuration comme MMCMM sélectionne la phrase centrale comme appartenant au discours de Mitterand.

3 Résultats sur le corpus d'apprentissage

Le corpus d'apprentissage comprenait 7523 phrases apprtenant au discours de Mitterand. L'extraction a donné 6479 textes dont 5782 correctement trouvés. Soit une précision de 0.89, un rappel de 0.76 et un Fscore de 0.82.

4 Résultats sur le corpus de test

Sur le corpus de test le rappel s'est effondré : 0.15. La précision a faiblie dans une moindre proportion : 0.77. Une partie de ce phénomène vient du fait que la tâche demandée était plus axée sur le rhétorique ou la distribution du vocabulaire que sur la sémantique. Comme cette méthode essaie au contraire de s'affranchir de la rhétorique et du choix du vocabulaire le résultat va de soi. Nous pouvons illustrer ce phénomène dès le départ :

1ere phrase trouvée et suivantes:

> La quantité elle est facilement fournie, car la qualité c'est de l'exigence, et l'exigence par rapport à soi-même. Il n'y a pas de haute société, de progrès, il n'y a pas de grands pays qui n'ait à son propre égard une forte exigence. On ne fait rien dans la mollesse, dans la faiblesse, et en tout cas certainement pas dans l'ignorance ! Les gouvernements antérieurs ont signé ce pacte qu'on appelle par ses initiales - GATT - cet accord commercial qui implique que les produits de substitution américains pénètrent à l'intérieur de l'Europe sans subir de taxe. Ils viennent donc concurrencer indûment nos propres productions. Si l'on ajoute à cela les montants compensatoires monétaires qui font que des pays comme la Hollande et l'Allemagne peuvent vendre en France des produits moins chers, alors que faire ? Et je vois des foules paysannes dressées pour dénoncer, pour maudire, les produits de substitution américains. Vous êtes ici affrontés, dans des conditions plus rigoureuses qu'ailleurs, à tous les problèmes urbains des grandes concentrations humaines. Il y a votre population, celle qui figure sur les documents officiels, 44000 habitants peut-être, puis il y a les autres, souvent les laissés-pour-compte ou les migrants, qui vont d'une ville à l'autre, ne sachant où se loger, où s'arrêter. Normalement attirés par ces lieux où l'on peut se perdre, tant et tant d'hommes cherchent à être reconnus, tant d'autres cherchent à ne pas l'être. Il vous faut assurer la synthèse de ces aspirations, de ces besoins, de ces moyens. Croyez-moi, la France c'est comme cela. La France c'est le pays que l'on sait, on connaît sa beauté, ses attraits, on connaît aussi ses misères. Depuis toujours il a été composé d'alluvions venues d'un peu partout au gré des combats, des conquêtes ou bien de leur reflux, au gré aussi des aventures humaines qui conduisent plus naturellement les hommes à venir là où l'on se sent bien plutôt qu'ailleurs et on se sent généralement bien en France à condition bien entendu que la France sache recevoir et accueillir, qu'elle s'ouvre plutôt que de se fermer. Elle a reçu d'immenses bienfaits. Il m'est agréable de souligner que, grâce à la haute conscience que vous avez de vos devoirs, le Brésil peut vivre dans un Etat de droit. Vous avez derrière vous, un siècle et demi d'existence et vous avez pu dans ce temps-là élargir les compétences, adapter la jurisprudence à

> l'évolution du droit public et de la société. Bref, vous constituez en quelque sorte l'organe régulateur de la machine complexe de l'Etat.

phrase du discours de Mitterand les précédant :

> Je ne m'attarderai pas sur ce problème, mais je dirai un mot quand même de la Communauté européenne dans un instant. Comme vous savez, après avoir été visité cette exploitation, je me suis rendu aux Haras d'Aurillac. J'y ai rencontré les organisations régionales agricoles. Je me suis exprimé selon ma conviction et dit ce que je pensais de l'avenir de l'agriculture française, et auvergnate en particulier, en insistant sur le fait que pour assurer les mutations essentielles et préserver la compétitivité française, il fallait accepter un certain nombre de risques. La Communauté européenne, j'en suis tout à fait partisan. J'ai voté tous les accords européens. En 1957, j'étais un partisan fervent du Marché commun agricole. C'était une bataille difficile et je suis extrêmement touché de voir aujourd'hui avec quel empressement ceux qui ne l'ont pas voté me reprocheraient de ne pas suffisamment réussir l'-entreprise dont ils ne voulaient pas. Ce qui prouve qu'on peut changer d'opinion, ce qui est honorable, et même de devenir, c'est généralement le caractère des néophytes, des zélateurs enthousiastes, même à retardement. Cette Communauté des Dix doit devenir, je l'espère, et je ne négligerai rien pour cela après avoir pris les précautions élémentaires pour un marché loyal, la Communauté des Douze. Elle a décidé en effet de limiter la production laitière. Je pose aux agriculteurs et j'ai posé partout, y compris à ses dirigeants nationaux, la question suivante : est-ce que vous voulez de l'Europe ou est-ce que vous ne la voulez pas ? Si vous ne la voulez pas, vous êtes logiques. Si vous en voulez, on est Dix. On est Dix, il faut l'accord des Dix. Et oui, parce qu'il y a eu une évolution de la Communauté, une mauvaise évolution. On s'est éloigné des dispositions du Traité de Rome qui fixaient la possibilité de voter à l'unanimité dans des cas très stricts de préservation des souverainetés nationales. Et c'est à la demande de la France dans les années 60, que cette loi a été rompue et que l'on a adopté un compromis, dit compromis de Luxembourg qui, en fait, contraint à l'unanimité dans les dispositions qui ne le méritent pas, ce qui bloque le système. A la diète de Pologne, rappelez-vous, il fallait que tous les députés de Pologne votassent à l'unanimité les lois. Vous imaginez si cela était comme cela en France ! Heureusement qu'il y a une majorité ! Mais en demander davantage donnerait vraiment le champ un peu trop libre à ceux qui n'aiment pas les majorités. L'Europe a fait cela à la demande de la France. La France a eu beaucoup d'initiatives dans cette Europe. C'est très bien. J'ai déjà eu le plaisir de vous accueillir dans le passé, pas forcément les mêmes, mais un certain nombre d'entre vous, et je suis très heureux de l'occasion qui m'est donnée de vous recevoir à nouveau et donc de revoir certains d'entre vous et de faire connaissance des autres. Cela a une très grande force symbolique, ce titre de "meilleur ouvrier de France". Il évoque bien des valeurs fondamentales, et d'abord l'amour du métier ; ensuite un grand savoir-faire, une capacité d'expression, j'ai dit tout à l'heure de beauté, d'esthétique, la maîtrise de l'outil et la perfection technique qui ne peuvent se passer de la maîtrise de l'esprit. C'est aussi un concours qui récompense des femmes et des hommes qui travaillent souvent dans des techniques de pointe, dans les métiers nouveaux. Il n'y a pas que les métiers traditionnels, bien qu'il faille aussi les honorer, mais il faut suivre l'évolution de la technique, l'évolution des

> temps ; il faut que la France dispose des meilleurs ouvriers possibles dans tous les domaines : ceux qu'on a coutume de connaître à travers les générations et ceux qui se révèlent comme des techniques et savoir-faire indispensables avec l'évolution de la technologie. Je félicite donc les lauréats, je ne pourrai tous les connaître, mais je suis heureux et flatté qu'ils soient aujourd'hui les hôtes de la Présidence de la République et je veux qu'ils rapportent chez eux, quand ils rentreront à la maison, dans leur famille, le sentiment d'avoir reçu le juste prix qui les honore et nous honore. Vous savez on ne se passera jamais de la qualité. La quantité elle est facilement fournie, car la qualité c'est de l'exigence, et l'exigence par rapport à soi-même.

Conclusion:

> Comme on peut le constater il s'agit avant tout de politique européenne. Par rapport à la séparation de thème politique intérieure / politique étrangère cette partie se situerait plutôt du coté de la politique intérieure.

Mais, dans le corpus d'apprentissage nous trouvons entre autre :

> <107:13:C> Vous avez également consacré beaucoup de temps et d'énergie à la construction de l'Europe de la Défense.
> <107:14:C> C'était et c'est encore un défi essentiel pour l'avenir de notre continent.
> <107:15:C> Cette Europe de la Défense, vous l'avez aidée à naître en surmontant tous les obstacles, en déployant des efforts constants.
> <107:16:C> Vous avez accompagné avec confiance et avec foi l'élan de Saint-Malo.
> <107:17:C> Vous avez su tirer pour nos armées toutes les conséquences du Conseil européen de Nice.
> <107:18:C> Si l'Europe de la Défense est désormais plus qu'une espérance, si elle est aujourd'hui une réalité qui s'enracine, c'est en partie à vous que la France et ses partenaires le doivent.
> <107:19:C> Il me revient enfin, en tant que chef des armées, de rendre hommage à vos mérites personnels.
> <107:20:C> Votre sens du devoir, votre franchise, votre fidélité sans faille à la mission sont d'abord des qualités de soldat.
> <107:21:C> Votre souci constant de la reconnaissance par la Nation du dévouement de nos armées et votre attachement à l'amélioration de la condition militaire vous ont valu le respect de tous.

Donc le fait de retrouver ces phrases associées à la politique européenne associées au discours de Jacques Chirac implique que cette partie ne pas doit être associée au discours de François Mitterand. Une lecture aléatoire de parties du corpus d'apprentissage montre que ce phénomène se reproduit de temps en temps. Les vecteurs de références s'en trouvent nécessairement affectés. Comme dans le corpus d'apprentissage le nombre de phrases associées aux discours de Jacques Chirac est beaucoup plus important, il est normal que les phrases traitant de la politique européenne se trouvent associées à ces discours. Bien sûr il existe peut-être d'autres facteurs que nous essaierons de dégager. La méthode présentée sera mieu adaptée à un traitement de classification en fonction du thème traité.

Références

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

Crestan E. , El-Bèze M. , de Loupy Claude (2003). Peut-on trouver la taille de contexte optimale en désambiguisation sémantique ? *Actes de TALN2003*11-14 juin, Batz-sur-Mer. Vol 1. Pp 85-94.

Ellman J., Tait, J. (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 Ð 217.

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

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

Yarowsky D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proc. of COLING92.*