LSA : Les Limites d'une Approche Statistique

Mathieu Roche, Jacques Chauché

▶ **To cite this version:**

Mathieu Roche, Jacques Chauché. LSA : Les Limites d'une Approche Statistique. 2006, pp.95-106. lirmm-00102797

HAL Id: lirmm-00102797
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00102797

Submitted on 2 Oct 2006

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.

LSA : les limites d'une approche statistique

Mathieu Roche et Jacques Chauché

Équipe TAL, LIRMM - UMR 5506, Université Montpellier 2,
34392 Montpellier Cedex 5 - France
{mroche,chauche}@lirmm.fr

Résumé. Cet article propose une méthode de classification conceptuelle à partir de textes qui sont par nature des données complexes. Nous nous sommes intéressés à la méthode statistique appelée LSA (Latent Semantic Analysis) utilisée pour regrouper des termes et/ou des textes. Cet article met particulièrement en relief les limites de LSA sur des données réelles. Des propositions pour améliorer la qualité des résultats sont enfin proposées.

1 Introduction

Cet article s'intéresse au regroupement des termes extraits à partir de corpus spécialisés. Nous définissons un terme comme un groupe de mots ayant des propriétés syntaxiques (de type Nom-Nom, Nom-Adjectif, etc.) et qui représente une trace linguistique de concepts (Kodratoff (2004)). Les concepts sont définis par l'expert du domaine. Par exemple, les termes *génie logiciel* et *intelligence artificielle* pourraient être associés au concept de *Cours en informatique*. Dans cet article, nous ne détaillerons pas la méthode d'extraction de la terminologie qui a été utilisée (voir l'approche décrite dans les travaux de Roche (2004)).

Plusieurs méthodes de classification de termes à partir de textes existent dans la littérature. La plupart de ces systèmes sont fondés sur des méthodes mixtes : linguistiques et statistiques. Par exemple, le système LEXICLASS (Assadi (1997, 1998)) utilise des mesures de similarité pour regrouper les termes partageant souvent un même contexte (par exemple, les termes qui sont souvent en présence du même adjectif peuvent être regroupés). Le système ASIUM développé par Faure et Nédellec (1998) utilise une hypothèse similaire fondée sur le fait que le contexte permet de déterminer la sémantique. Le système ASIUM qui possède une approche coopérative avec l'expert utilise les connaissances syntaxiques (obtenues avec un analyseur syntaxique) et des mesures de similarité pour construire une classification conceptuelle. Le système ROWAN (Fontaine et Kodratoff (2002)) construit des classes de termes et de relations syntaxiques (Sujet-Verbe, Verbe-Objet, etc.). Outre l'approche coopérative avec l'expert de ROWAN (Fontaine et Kodratoff (2002)), un algorithme d'induction a également été proposé par Kodratoff (2004). Notons qu'un résumé de l'état de l'art des méthodes de classification de termes à partir de textes est présenté dans l'article de Aussenac-Gilles et Bourigault (2003).

Comme nous l'avons précisé, la plupart des systèmes de classification conceptuelle à partir de textes utilisent des approches mixtes. Dans cet article, nous allons nous appuyer sur la méthode appelée Latent Semantic Analysis (LSA) développée par Landauer et Dumais (1997);

Landauer et al. (1998)[1]. LSA est une méthode uniquement fondée sur une approche statistique appliquée à des corpus de grande dimension consistant à regrouper les termes (classification conceptuelle) ou les contextes (classification de textes).

Cet article présente la méthode LSA en mettant particulièrement en relief les limites de celle-ci. Ces limites sont dues à la complexité des données textuelles. Ainsi, nous souhaitons ici discuter des résultats obtenus avec LSA qui peuvent se révéler parfois décevants. Nous tenterons d'apporter des hypothèses afin d'expliquer et de discuter de ces limites. Enfin, nous proposerons des pistes de travail que nous souhaitons mettre en œuvre dans l'équipe TAL du LIRMM pour améliorer la qualité des résultats obtenus avec LSA. L'approche que nous souhaitons proposer consiste à apporter des informations syntaxiques à LSA.

2 Latent Semantic Analysis (LSA)

La méthode LSA qui s'appuie sur l'hypothèse « harrissienne », est fondée sur le fait que des mots qui apparaissent dans le même contexte sont sémantiquement proches. Le corpus est représenté sous forme matricielle. Les lignes sont relatives aux mots et les colonnes représentent les différents contextes choisis (un document, un paragraphe, une phrase, etc.). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le cosinus de l'angle entre les deux vecteurs.

Caractéristiques théoriques de LSA

La théorie sur laquelle s'appuie LSA est la décomposition en valeurs singulières (SVD). Une matrice $A = [a_{ij}]$ où a_{ij} est la fréquence d'apparition du mot i dans le contexte j, se décompose en un produit de trois matrices USV^T. U et V sont des matrices orthogonales et S une matrice diagonale. La figure 1 représente le schéma bien connu d'une telle décomposition où r représente le rang de la matrice A.

Soit S_k où $k < r$ la matrice produite en enlevant de S les $r - k$ colonnes qui ont les plus petites valeurs singulières. Soit U_k et V_k les matrices obtenues en enlevant les colonnes correspondantes des matrices U et V. La matrice $U_k S_k V_k^T$ peut alors être considérée comme une version compressée de la matrice originale A.

Il est coutume de dire que LSA est une méthode statistique ou numérique car elle s'appuie sur une théorie mathématique bien connue. Cependant, on peut également dire que LSA est une méthode géométrique car seuls des résultats d'algèbre linéaire sont utilisés.

Nous précisons qu'avant d'effectuer la décomposition en valeurs singulières, nous effectuons une première étape de normalisation de la matrice d'origine A. Cette normalisation consiste à appliquer un logarithme et un calcul d'entropie sur la matrice A. Ainsi, plutôt que de se fonder directement sur le nombre d'occurrences de chacun des mots, une telle transformation permet de s'appuyer sur une estimation de l'importance de chacun des mots dans leur

[1] voir aussi, http ://www.msci.memphis.edu/∼wiemerhp/trg/lsa-followup.html

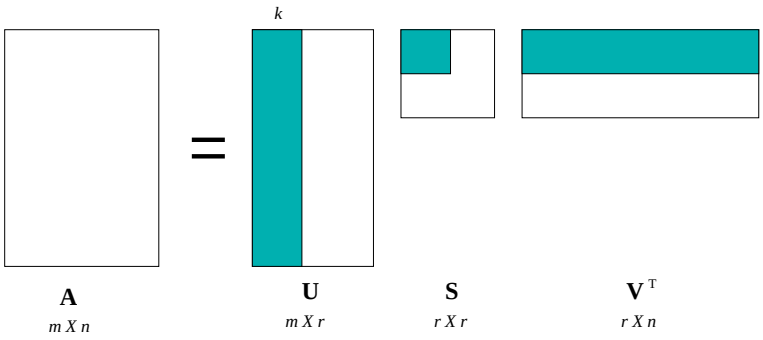

FIG. 1 – *Décomposition en valeurs singulières. La matrice A représente le corpus d'origine de m lignes (mots du corpus) et n colonnes (contextes).*

contexte. De manière similaire aux travaux de Turney (2001), cette étape de normalisation peut également s'appuyer sur la méthode du tf×idf, approche bien connue dans le domaine de la Recherche d'Information (Salton (1991)).

Précisons de plus que nous ne prenons pas en compte les ponctuations ainsi qu'un certain nombre de mots non significatifs du point de vue sémantique tels que les mots « et », « à », « le », etc. L'utilisation d'une telle liste de mots « vides » influence positivement le résultat final (Roche et Kodratoff (2003)).

Les expériences décrites dans la section suivante ont été menées avec un nombre de facteurs k égal à 200. Comme le montrent Landauer et Dumais (1997) le choix du nombre de facteurs de la matrice est un paramètre qui influence sensiblement le résultat final. Nous avons expérimenté différents paramètres, et comme dans les expériences de Wiemer-Hastings (2000), nous avons estimé qu'un nombre de facteurs égal à 200 donnait les meilleurs résultats sur notre corpus qui est décrit dans la section suivante.

3 Expérimentations

3.1 Caractéristiques du corpus étudié

Dans cet article, nous allons nous appuyer sur un corpus des Ressources Humaines de la société PerformanSe écrit en français[2]. Une caractéristique essentielle de ce corpus, d'une taille de 3784 Ko, est qu'il utilise un vocabulaire spécialisé comme la plupart des corpus que nous étudions. Cependant, une spécificité du corpus des Ressources Humaines est qu'il contient des tournures de phrases revenant souvent, ce qui peut influencer positivement le traitement avec LSA. À titre d'exemple, la meilleure similarité trouvée entre deux termes a été obtenue grâce à deux phrases ayant des tournures strictement identiques et pour lesquelles seul un terme différait. Dans nos expérimentations sur le corpus des Ressources Humaines, chaque phrase

[2]Fragment du corpus disponible à l'adresse : http ://www.lri.fr/∼roche/Recherche/corpusPsy.html

représente les contextes, c-à-d. les colonnes dans notre matrice relative au corpus. Nous précisons que pour d'autres corpus explicitement découpés en documents distincts (par exemple, des corpus de résumés), les contextes peuvent être représentés par les documents eux-mêmes.

3.2 Protocole expérimental

Afin d'effectuer des regroupements de termes, il est nécessaire d'obtenir des similarités entre les termes de bonne qualité deux à deux. Ainsi, dans les expérimentations que nous allons décrire, nous nous intéressons aux couples de termes trouvés automatiquement par LSA. Nous allons nous appuyer sur le corpus des Ressources Humaines pour lequel plus de 1800 termes présents ont été extraits dans le corpus à l'aide de la méthode d'extraction de la terminologie décrite dans les travaux de Roche (2004). Ces termes ont alors été associés manuellement à un concept par un expert du domaine. Par exemple, avec ce corpus, l'expert a défini le concept « Relationnel » dont les termes *confrontation-ouverte*, *contact-superficiel* et *entourage-compréhensif* sont des instances.

Afin d'évaluer la qualité des regroupements donnés par LSA, nous allons effectuer deux types de mesures. La première mesure correspond au pourcentage de couples corrects (c-à-d., les termes associés par LSA qui appartiennent à un même concept). Dans le domaine de l'apprentissage, une telle mesure correspond à la Précision. La deuxième mesure que nous allons utiliser est la couverture des termes classés par LSA. Cette mesure d'évaluation consiste à calculer le pourcentage de termes de la classification conceptuelle présents dans les couples formés par LSA.

4 Première limite majeure de LSA : l'influence de la taille des contextes

Nous avons alors effectué différentes expérimentations (voir TAB. 1) sur le corpus lemmatisé des Ressources Humaines. En effet, nous avons montré que le corpus lemmatisé donnait de meilleurs résultats que le corpus non lemmatisé (Roche et Kodratoff (2003)).

Les expérimentations effectuées sur ce corpus montrent que le nombre de couples de termes corrects, c-à-d. appartenant à un même concept, est globalement faible. Dans TAB. 1, nous pouvons également noter que plus la similarité entre les termes est proche de 1 (similarité maximale) et plus la qualité des couples de termes extraits est intéressante. En effet, en augmentant le seuil de similarité de 0.1, le pourcentage de couples de termes corrects augmente de plus de 10%.

Bestgen (2004) précise que la taille des contextes (documents) est primordiale pour obtenir une qualité des résultats satisfaisante. Cette affirmation confirme les travaux de Rehder et al. (1998) qui ont effectué des expérimentations pour estimer la taille minimale d'un contexte afin d'obtenir des résultats intéressants avec LSA. Ces expérimentations ont consisté à découper les documents d'un corpus correspondant à des essais d'étudiants en documents de 10 mots, 20 mots, et ceci jusqu'à 200 mots. Les expérimentations ont montré que si les contextes (do-

Similarité (cosinus)	0.3	0.4	0.5	0.6
% de termes correctement associés (c-à-d. % de couples corrects)	**19.2 %** (31/161)	**32.1 %** (9/28)	**42.9 %** (3/7)	**75.0%** (3/4)
% de termes de la classification manuelle présents dans les couples	**9.8 %** (180/1842)	**2.7 %** (49/1842)	**0.8 %** (14/1842)	**0.4 %** (8/1842)

TAB. 1 – *Expérimentations sur le corpus des Ressources Humaines de la société PerformanSe.*

cuments) possèdent moins de 60 mots alors la méthode LSA se révèle décevante.

Les contextes de notre corpus correspondent aux phrases qui sont composées de beaucoup moins de 60 mots. En effet, notre corpus possède, en moyenne, 27 mots par phrase. Cette situation pourrait donc expliquer les résultats pas toujours convaincants que nous avons obtenus avec LSA.

Notre motivation d'utiliser LSA repose sur une représentation mathématique solide des textes qui est indépendante des langues. La méthode LSA de base est automatique sans la nécessité de s'appuyer sur des connaissances linguistiques et du domaine (dictionnaires, ontologies, grammaires, étiquettes morpho-syntaxiques, etc.). Ceci est un avantage car la méthode LSA est largement généralisable selon les langues et les domaines des textes de spécialité étudiés. Mais cela peut aussi être une faiblesse de l'approche lorsque l'on a des objectifs pointus à atteindre comme nous allons le présenter dans la section suivante (section 5). La section 6 montrera enfin que l'ajout de connaissances syntaxiques à la méthode générale et automatique LSA pourrait améliorer les résultats.

5 Seconde limite majeure de LSA : difficulté dans le cas de la proximité du vocabulaire utilisé

Dans cette section, nous allons définir une autre tâche à effectuer prenant en compte des contextes de taille plus importante. Ainsi, nous allons mener des expérimentations consistant à classifier des textes. Pour cela, nous allons travailler à partir d'un autre corpus explicitement divisé en différentes thématiques. Nous proposons de travailler avec le corpus issu de la tâche Novelty de la compétition internationale TREC 2004 à laquelle l'axe *Fouille de Textes* de l'équipe IA[3] du LRI[4] mené par Yves Kodratoff a participé (Amrani et al. (2004)). Ce corpus est composé d'articles journalistiques divisés en cinquante thématiques distinctes.

Le but de la tâche Novelty était de retrouver les phrases pertinentes et nouvelles dans les articles journalistiques. Pour chacun des cinquante thèmes traités dans les articles du corpus, une description explicite de la pertinence et de la nouveauté était donnée aux participants de TREC Novelty. Lors de l'édition 2004 de TREC Novelty, des textes non pertinents ont pu être

[3]http ://www.lri.fr/ia/
[4]http ://www.lri.fr/

rajoutés dans les différentes thématiques.

Dans cet article, nous nous sommes plus spécifiquement intéressés à un sous-ensemble du corpus normalisé (Amrani et al. (2004)) rassemblant 29 articles journalistiques traitant d'une opération sur la greffe de la main. Les textes pertinents étaient ceux relatifs à la première greffe de la main effectuée sur le patient Matthew David Scott. Cependant, dans ce corpus, quatre articles se révèlent non pertinents (textes du domaine médical mais ne décrivant pas explicitement l'opération de Matthew David Scott).

Ainsi, nous avons utilisé LSA dans le but de vérifier que ces quatre textes avaient un score faible (le plus éloigné possible de 1) comparativement à l'ensemble des textes du corpus. De manière globale, ce résultat attendu n'a pas été vérifié dans nos expérimentations (voir TAB. 2). Plus précisément, les résultats montrent que deux des textes non pertinents ont un score de similarité plus faible que le score moyen (0.0933) mais les deux autres articles non pertinents ont un score plus important que ce score moyen. Ainsi, il semble globalement difficile de conclure sur l'efficacité de la méthode dans ce cas. Ceci peut s'expliquer par le fait que le vocabulaire utilisé dans les textes pertinents et non pertinents est souvent très proche. Ainsi, une classification de textes très fine peut se révéler difficile avec LSA.

Numéro des articles	Moyenne des scores	Numéro des articles	Moyenne des scores	Numéro des articles	Moyenne des scores
1 *(np)*	0.0357	11	0.0972	21	0.0970
2 *(np)*	0.0365	12	0.0922	22	0.1135
3 *(np)*	0.1083	13	0.0929	23	0.0990
4 *(np)*	0.1006	14	0.0961	24	0.1077
5	0.1068	15	0.0594	25	0.1208
6	0.1347	16	0.0951	26	0.0639
7	0.1191	17	0.0985	27	0.0395
8	0.1042	18	0.0958	28	0.1016
9	0.0952	19	0.1085	29	0.0354
10	0.1311	20	0.1197	**Moyenne des 29 textes : 0.0933**	

TAB. 2 – *Pour chaque article, la moyenne des scores de similarité avec les 28 autres articles est calculée. La notation "np" désigne les articles non pertinents.*

Dans un second temps, nous avons ajouté des textes issus d'une autre thématique pour étudier la similarité entre ces textes clairement non pertinents par rapport au sous-corpus traitant de la greffe de la main. Par exemple, TAB 3 montre que les quatre textes de thématiques singulièrement différentes[5] choisis aléatoirement ont un score de similarité globalement plus faible que les scores moyens. Ceci s'explique par le vocabulaire de ces quatre articles qui est éloigné des articles relatifs à la greffe de la main.

[5]Deux articles traitent du procès de Pinochet et deux articles sont relatifs au premier commandement d'une navette spatiale par une femme, le commandant Eileen Collins.

Sous-corpus original de "la greffe de la main" (avec les 4 premiers articles non pertinents)	
Numéro des articles	Moyenne des scores
1 *(np)*	0.0338
...	...
28	0.0966
29	0.0332
30 *(np)*	0.0482
31 *(np)*	0.0444
32 *(np)*	0.0393
33 *(np)*	0.0324
Moyenne des 33 textes : 0.0826	

Sous-corpus modifié de "la greffe de la main" (sans les 4 premiers articles non pertinents)	
Numéro des articles	Moyenne des scores
5	0.0992
...	...
28	0.0993
29	0.0318
30 *(np)*	0.0505
31 *(np)*	0.0469
32 *(np)*	0.0422
33 *(np)*	0.0353
Moyenne des 29 textes : 0.0873	

TAB. 3 – *Pour chaque article non pertinent (numéros 30 à 33) ajouté au sous-corpus relatif à la greffe de la main, la moyenne des scores de similarité avec l'ensemble des autres articles est calculée. La notation "np" désigne les articles non pertinents.*

6 Solution proposée : Ajout de connaissances syntaxiques à LSA

Afin d'améliorer la performance de LSA, Wiemer-Hastings (2000) propose d'ajouter des connaissances syntaxiques à LSA en transformant les phrases en structures syntaxiques. Pour ce faire, une segmentation syntaxique des phrases en trois groupes de mots est effectuée :

- syntagmes nominaux représentant les sujets,
- verbes en prenant en compte les adverbes et les syntagmes adverbiaux,
- syntagmes nominaux représentant les objets.

Ainsi, chaque phrase est représentée sous la forme (« verbe » « sujet » « objet »). Lorsqu'il y a deux objets (« objet1 » et « objet2 ») affectés à un même verbe, la phrase sera représentée sous la forme (« verbe » « sujet » « objet1 ») et (« verbe » « sujet » « objet2 »), de même dans le cas de la présence de deux sujets associés à un seul verbe.

Initialement, LSA ne prend pas en compte un certain nombre de mots (« stops words ») tels que « *if* », « *because* », « *have* », etc. Contrairement à la version originale de LSA, Wiemer-Hastings (2000) prend en compte de tels mots et peut les utiliser pour construire les structures de certaines phrases. Par exemple, la phrase *if the new motherboard uses the same type of RAM* sera représentée sous la forme *(« if uses » « the new motherboard » « the same type of RAM »)*.

Les résultats expérimentaux développés dans l'étude de Wiemer-Hastings (2000) restent malgré tout décevants. Ceci pourrait être dû à une analyse syntaxique pas assez fine. De plus, les données réelles utilisées (évaluation de la qualité des réponses d'étudiants) dans les travaux

de Wiemer-Hastings (2000) pouvaient se révéler trop difficiles à traiter [6].

Dans les futurs travaux que nous souhaitons mener dans l'équipe TAL du LIRMM, nous proposons également d'ajouter des connaissances syntaxiques à LSA. Pour cela, nous allons utiliser l'analyseur syntaxique SYGMART développé par Chauché (1984). À partir de textes bruts écrits en français, SYGMART fournit une analyse sous forme d'arbres morpho-syntaxiques. La qualité de l'analyseur syntaxique utilisé pour le français ainsi qu'une analyse plus fine des phrases devrait nous permettre d'améliorer les résultats obtenus par LSA.

Pour cela, nous allons décomposer chaque phrase de la manière la plus fine possible en utilisant au mieux la précision de l'analyseur SYGMART. Pour cela, nous allons nous appuyer sur la décomposition proposée par Wiemer-Hastings (2000) en trois entités : *sujet*, *verbe* et *objet*. De plus, nous pouvons ajouter toutes les informations syntaxiques supplémentaires apportées par SYGMART. Pour chacun des éléments, seuls les gouverneurs des syntagmes nominaux sont conservés. Illustrons une telle décomposition avec deux exemples donnés ci-dessous dans lesquels les termes *projet-de-recherche* et *ambitieuses-perspectives* sont déterminés par l'expert[7]. Ces deux exemples de phrases possèdent seulement deux mots en communs (*ajout* et *connaissance*). De ce fait, la méthode LSA pourrait avoir tendance à donner une similarité décevante entre les termes *projet-de-recherche* et *ambitieuses-perspectives*.

EXEMPLE 1
Phrase
L'ajout de connaissances syntaxiques à la méthode statistique LSA caractérise notre projet-de-recherche à moyen-terme
Décomposition
sujet*(ajout, connnaissance*, complément*(méthode, LSA))* **verbe***(caractériser)* **objet***(projet-de-recherche,* complément*(moyen-terme))*

EXEMPLE 2
Phrase
L'ajout de connaissances sémantiques significatives à notre approche ouvre également d'ambitieuses-perspectives
Décomposition
sujet*(ajout, connnaissance*, complément*(approche)* **verbe***(ouvrir)* **objet***(ambitieuses-perspectives)*

[6] Les textes traités dans Wiemer-Hastings (2000) ont des contextes de taille très réduite. Avec les phrases ayant un contexte moyen de 16 mots du corpus traité par Wiemer-Hastings (2000), les résultats qui sont obtenus peuvent tout de même se révéler comparables aux performances humaines. Cependant, comme le précisent Wiemer-Hastings et al. (1999), il semble difficile même pour un expert humain d'obtenir de bonnes performances avec des contextes ayant peu de mots.

[7] Un trait d'union a été placé entre les mots composant les termes pour que ces derniers soient reconnus comme des mots à part entière par les analyseurs syntaxiques.

En utilisant, les connaissances syntaxiques, les deux phrases possèdent exactement les deux mêmes mots principaux pour caractériser le sujet. Ainsi, il pourrait être intéressant de privilégier la proximité sémantique des deux termes présents comme objets des deux phrases. Une manière de prendre en compte cette information syntaxique dans la méthode LSA pourrait, par exemple, consister à ajouter une valeur numérique (notée α) au score trouvé par LSA entre les deux termes partageant le même contexte (ici, les termes *projet-de-recherche* et *ambitieuses-perspectives* partagent le même sujet). Par exemple, α pourrait prendre comme valeur le maximum ou la moyenne des scores obtenus avec LSA.

Nous avons effectué des premières expérimentations avec les deux phrases de l'exemple cité ci-dessus. Dans ces expérimentations, nous avons souhaité évaluer le taux de similarité obtenu avec LSA entre ces deux phrases comparativement à l'ensemble des phrases de l'introduction de ce présent article (section 1). Nous avons alors ajouté la deuxième phrase de notre exemple (phrase numérotée 2) aux 22 phrases représentant l'introduction (phrases numérotées de 3 à 24). Nous avons évalué le taux de similarité de la première phrase de l'exemple (phrase numérotée 1) avec les 23 phrases constituant le corpus. Ainsi, les deux meilleurs taux de similarité (avec des valeurs numériques du même ordre) avec la première phrase de l'exemple (phrase 1) ont été obtenus avec les phrases 2 et 7 (voir ci-dessous).

Phrase 1
L'ajout de connaissances syntaxiques à la méthode statistique LSA caractérise notre projet-de-recherche à moyen-terme
Phrase 2
L'ajout de connaissances sémantiques significatives à notre approche ouvre également d'ambitieuses-perspectives
Phrase 7
Dans cet article, nous ne détaillerons pas la méthode d'extraction de la terminologie qui a été utilisée (voir l'approche décrite dans les travaux de Roche).

La méthode que nous souhaitons mettre en œuvre privilégiera la similarité des deux premières phrases qui ont les mêmes mots principaux pour caractériser le sujet. Pour cela, nous donnerons une valeur de similarité plus importante au couple de phrases (1,2) comparativement au couple (1,7). Ces deux couples de phrases avaient initialement un score de similarité du même ordre obtenu avec LSA[8]. Ainsi, cette méthode permettra de privilégier la similarité entre les termes *projet-de-recherche* et *ambitieuses-perspectives* issus des phrases 1 et 2.

Les travaux de Wiemer-Hastings (2000) mais également l'approche de Faure et Nédellec (1998) s'appuient essentiellement sur les verbes pour effectuer un regroupement de termes. Par exemple, pour construire des classes sémantiques, le système ASIUM (Faure et Nédellec (1998); Faure (2000)) utilise la notion de « proximité sémantique » fondée sur le principe de distance entre les mots qui partagent le même contexte (en particulier, les verbes ayant souvent les mêmes objets). Les contextes "verbe objet" ou "sujet verbe" peuvent en effet être davantage pertinents pour effectuer un regroupement de termes comparativement à une relation "sujet objet" (sans considérer les verbes). Ainsi, nous pourrions utiliser un poids (α) à ajouter au score

[8]Dans ces expérimentations, le corpus constitué étant de taille très réduite, nous avons utilisé l'application http ://lsa.colorado.edu/cgi-bin/LSA-one2many.html avec le domaine "Français-Total".

de LSA plus important pour les termes trouvés dans les relations "verbe objet" ou "sujet verbe" partageant le même verbe comparativement à une relation "sujet objet" partageant un même contexte (sujet ou objet). Un processus d'apprentissage supervisé pourrait bien entendu être mis en œuvre pour établir les poids les plus adaptés à appliquer (voir l'ouvrage de Cornuéjols et al. (2002) qui présente un état de l'art des différentes méthodes d'apprentissage).

Nous avons présenté ici les perspectives à moyen terme que nous souhaitons mettre en place et qui, nous l'espérons, permettront d'améliorer les résultats parfois décevants obtenus en utilisant la méthode LSA de base.

7 Conclusion et perspectives

La méthode LSA qui est une méthode statistique (ou géométrique) ne prend pas en compte l'ordre des mots. Par exemple, les phrases *"le mot français est écrit dans le corpus"* et *"le corpus est écrit avec des mots français"* ont un sens très différent. Cependant, la méthode LSA conclura à une similarité parfaite entre ces deux phrases qui partagent les mêmes mots (sans considérer les mots « vides »).

Le fait d'ajouter des connaissances syntaxiques permet d'acquérir une meilleure compréhension du sens et donc de construire une classification conceptuelle de meilleure qualité. Ceci est une piste de travail particulièrement intéressante à développer. En effet, comme nous l'avons montré dans cet article, la méthode LSA possède des limites. Ainsi, l'ajout de connaissances syntaxiques pourrait pallier aux limites de LSA décrites dans cet article.

La phrase relative au deuxième exemple de la section 6 de cet article illustre une autre piste de travail... La perspective énoncée dans cet exemple propose d'ajouter des connaissances sémantiques à la méthode LSA. Ceci pourrait consister à remplacer des mots par le nom d'un concept plus général. Ceci permettrait également d'améliorer les résultats obtenus avec LSA.

Remerciements

Les auteurs remercient Yves Kodratoff (LRI) et Serge Baquedano (Société PerformanSe) pour le travail effectué sur le corpus des Ressources Humaines.

Références

Amrani, A., J. Azé, T. Heitz, Y. Kodratoff, et M. Roche (2004). From the texts to the concepts they contain : a chain of linguistic treatments. In *In Proceedings of TREC'04 (Text REtrieval Conference)*, pp. 712–722.

Assadi, H. (1997). Knowledge acquisition from texts : Using an automatic clustering method based on noun-modifier method. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 504–509.

Assadi, H. (1998). *Construction d'ontologies à partir de textes techniques - application aux systèmes documentaires*. Ph. D. thesis, Université de Paris 6.

Aussenac-Gilles, N. et D. Bourigault (2003). Construction d'ontologies à partir de textes. In *Actes de TALN03*, Volume 2, pp. 27–47.

Bestgen, Y. (2004). Analyse sémantique latente et segmentation automatique de textes. In *Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data)*, Volume 1, pp. 171–181.

Chauché, J. (1984). Un outil multidimensionnel de l'analyse du discours. In *In Proceedings of Coling, Standford University, California*, pp. 11–15.

Cornuéjols, A., L. Miclet, et Y. Kodratoff (2002). *Apprentissage artificiel, Concepts et algorithmes*. Eyrolles.

Faure, D. (2000). *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph. D. thesis, Université Paris-Sud.

Faure, D. et C. Nédellec (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In P. Velardi (Ed.), *LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications*, Granada Espagne, pp. 5–12.

Fontaine, L. et Y. Kodratoff (2002). Comparaison du rôle de la progression thématique et de la texture conceptuelle chez les scientifiques anglophones et francophones s'exprimant en anglais. In *Actes de la Journée de Rédactologie scientifique : L'écriture de la recherche*.

Kodratoff, Y. (2004). Induction extensionnelle : définition et application l'acquisition de concepts à partir de textes. *Revue RNTI E2, numéro spécial EGC'04 1*, 247–252.

Landauer, T. et S. Dumais (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review 104*(2), 211–240.

Landauer, T. K., P. W. Foltz, et D. Laham (1998). Introduction to latent semantic analysis. In *Discourse Processes*, Volume 25, pp. 259–284.

Rehder, B., M. Schreiner, M. Wolfe, D. Laham, T. Landauer, et W. Kintsch (1998). Using latent semantic analysis to assess knowledge : Some technical considerations. In *Discourse Processes*, Volume 25, pp. 337–354.

Roche, M. (2004). *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. Ph. D. thesis, Université de Paris 11.

Roche, M. et Y. Kodratoff (2003). Utilisation de LSA comme première étape pour la classification des termes d'un corpus spécialisé. In *Actes (CD-ROM) de la conférence MAJECSTIC'03 (MAnifestation des JEunes Chercheurs dans le domaine STIC)*.

Salton, G. (1991). Developments in automatic text retrieval. *Science 253*, 974–979.

Turney, P. (2001). Mining the Web for synonyms : PMI–IR versus LSA on TOEFL. In *Proceedings of ECML'01, Lecture Notes in Computer Science*, pp. 491–502.

Wiemer-Hastings, P. (2000). Adding syntactic information to LSA. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, pp. 989–993.

Wiemer-Hastings, P., K. Wiemer-Hastings, et A. Graesser (1999). Improving an intelligent

tutor's comprehension of students with Latent Semantic Analysis. *Artificial Intelligence in Education*, 535–542.

Summary

This paper proposes a clustering method from texts which are complex data. We interested in statistical method called LSA (Latent Semantic Analysis) used to the terms and/or texts clustering. This paper shows the limits of LSA on real data. Finally, we explain how we can improve the results.