ExpLSA : utilisation d'informations syntaxico-sémantiques associées à LSA pour améliorer les méthodes de classification conceptuelle

Nicolas Béchet, Mathieu Roche, Jacques Chauché

$ExpLSA$: utilisation d'informations syntaxico-sémantiques associées à LSA pour améliorer les méthodes de classification conceptuelle

Nicolas Béchet, Mathieu Roche, Jacques Chauché

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

Résumé. L'analyse sémantique latente (LSA - Latent Semantic Analysis) est aujourd'hui utilisée dans de nombreux domaines comme la modélisation cognitive, les applications éducatives mais aussi pour la classification. L'approche présentée dans cet article consiste à ajouter des informations grammaticales à LSA. Différentes méthodes pour exploiter ces informations grammaticales sont étudiées dans le cadre d'une tâche de classification conceptuelle.

1 Introduction

Le domaine de la classification de données textuelles se décline en de nombreux axes parmi lesquels la classification conceptuelle. Cette dernière consiste à regrouper des termes dans des concepts définis par un expert. Citons par exemple les termes *pot d'échappement*, *pare-brise* et *essuie glace* qui peuvent être classés dans le concept *automobile*. Afin d'établir une telle classification sémantique, la proximité de chacun des termes issus des textes doit être mesurée. Ces termes sont ensuite classés en fonction de leurs proximités sémantiques par un algorithme de fouille de données tels que les *Kppv (K plus proches voisins)* ou bien les *K moyennes* (Cornuéjols et Miclet (2002)).

Nous nous focalisons dans cet article sur la première étape de la réalisation d'une classification conceptuelle : l'étude de la proximité des termes. Afin de calculer une telle proximité, nous nous appuyons sur une méthode appelée Latent Semantic Analysis (LSA) développée par Landauer et Dumais (1997)[1]. La méthode LSA est 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). Une fois l'analyse sémantique latente appliquée à un corpus, un espace sémantique associant chaque mot à un vecteur est retourné. La proximité de deux mots peut alors être obtenue par un calcul de similarité comme le cosinus entre deux vecteurs. L'objectif de nos travaux est d'améliorer les performances de LSA par une approche nommée *ExpLSA* (***Exp***ansion des contextes avec ***LSA***).

L'approche *ExpLSA* consiste à enrichir le corpus qui constituera l'entrée d'une analyse sémantique latente *classique*. Cet enrichissement utilise les informations sémantiques obtenues

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

grâce à la syntaxe, ce qui permet d'utiliser *ExpLSA* aussi bien avec des corpus spécialisés ou non. Il n'est en effet pas utile d'utiliser un corpus d'apprentissage et donc pas nécessaire de connaître le thème général du corpus.

Dans cet article, nous allons nous appuyer sur un corpus des Ressources Humaines de la société PerformanSe[2] écrit en français[3]. Notons que les premiers travaux sur ce corpus ont été initiés dans l'équipe IA du LRI (Roche et Kodratoff (2003)). Une caractéristique essentielle de ce corpus est qu'il utilise un vocabulaire spécialisé. Par ailleurs, il contient des tournures de phrases revenant souvent, ce qui peut influencer positivement le traitement avec LSA. Ce corpus a fait l'objet d'une expertise manuelle nous permettant ainsi de valider nos expérimentations.

Nous proposons dans la section suivante de détailler les caractéristiques théoriques de la méthode LSA ainsi que les limites d'une telle analyse. La section 3 propose un état de l'art dans le domaine de l'utilisation de connaissances syntaxiques associées à LSA. Nous présentons ensuite notre méthode en y développant ses différentes étapes (section 4). Nous décrirons également (section 5) le protocole expérimental utilisé pour finalement présenter les résultats obtenus.

2 LSA

La méthode LSA qui s'appuie sur l'hypothèse "harrissienne" est fondée sur le fait que des mots qui apparaissent dans un 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.

2.1 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.

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. Les expériences décrites dans la section 5 ont été menées avec un nombre de facteurs k égal à 100, facteur faible qui est davantage approprié à des contextes de taille réduite.

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.

[2] http ://www.performanse.fr/

[3] Fragment du corpus disponible à l'adresse : http ://www.lirmm.fr/∼mroche/Recherche/corpusPsy.html

Nous précisons qu'avant d'effectuer la décomposition en valeurs singulières, une première étape de normalisation de la matrice d'origine A est exécutée. 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 contexte. De manière similaire aux travaux de Turney (2001), cette étape de normalisation peut également s'appuyer sur la méthode du tf$\times$idf, approche bien connue dans le domaine de la Recherche d'Information. 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.

2.2 Les limites de LSA

LSA offre des avantages parmi lesquels, la notion d'indépendance par rapport à la langue du corpus étudié, le fait de se dispenser de connaissances linguistiques ainsi que de celles du domaine tels que des thésaurus. Bien que cette approche soit prometteuse, il n'en demeure pas moins que son utilisation soulève des contraintes.

Notons tout d'abord l'importance de la taille des contextes choisis. Rehder et al. (1998) ont montré lors de leurs expérimentations que si les contextes possèdent moins de 60 mots, les résultats s'avèrent être décevants. Il a également été mis en évidence par Roche et Chauché (2006) que l'efficacité de LSA est fortement influencée par la proximité du vocabulaire utilisé.

Pour résoudre de tels problèmes, une des solutions peut consister à ajouter des connaissances syntaxiques à LSA, comme cela est décrit dans la section suivante.

3 État de l'art sur l'ajout de connaissances syntaxiques à LSA

Landauer et al. (1997) posent le problème du manque d'informations syntaxiques dans LSA en comparant cette méthode à une évaluation humaine. Il est question de proposer à des experts humains d'attribuer des notes à des essais sur le cœur humain de 250 mots rédigés par des étudiants. Un espace sémantique a été créé à partir de 27 articles écrits en anglais traitant du cœur humain "appris" par LSA. Les tests effectués concluent que la méthode LSA obtient des résultats satisfaisants comparativement à l'expertise humaine. Il en ressort que les mauvais résultats étaient dus à une absence de connaissances syntaxiques dans l'approche utilisée. Ainsi, les travaux qui sont décrits ci-dessous montrent de quelle manière de telles connaissances peuvent être ajoutées à LSA.

La première approche de Wiemer-Hastings et Zipitria (2001) utilise des étiquettes grammaticales (Brill (1994)) appliquées à l'ensemble du corpus étudié (corpus de textes d'étudiants). Les étiquettes étant rattachées à chaque mot avec un blanc souligné ("_"), l'analyse qui s'en suit via LSA considère le mot associé à son étiquette comme un seul terme. Les résultats de calculs de similarités obtenus avec une telle méthode restent décevants. Notons que de telles informations grammaticales ne sont pas des connaissances syntaxiques proprement dites contrairement à la seconde approche de Wiemer-Hastings et Zipitria (2001) décrite ci-dessous. Cette seconde approche se traduit par l'utilisation d'un analyseur syntaxique afin de segmenter le texte avant d'appliquer l'analyse sémantique latente. Cette approche est appelée "LSA structurée" (SLSA). Une décomposition syntaxique des phrases en différents composants (sujet,

verbe, objet) est tout d'abord effectuée. La similarité est ensuite calculée en traitant séparément par LSA les trois ensembles décrits précédemment. Les similarités (calcul du cosinus) entre les vecteurs des trois matrices formées sont alors évaluées. La moyenne des similarités est enfin calculée. Cette méthode a donné des résultats satisfaisants par rapport à "LSA classique" en augmentant la corrélation des scores obtenus avec les experts pour une tâche d'évaluation de réponses données par des étudiants à un test d'informatique.

Kanejiya et al. (2003) proposent un modèle appelé SELSA. Au lieu de générer une matrice de co-occurrences mot/document, il est proposé une matrice dans laquelle chaque ligne contient toutes les combinaisons mot_étiquette et en colonne les documents. L'étiquette "préfixe" renseigne sur le type grammatical du voisinage du mot traité. Le sens d'un mot est en effet donné par le voisinage grammatical duquel il est issu. Cette approche est assez similaire à l'utilisation des étiquettes de Brill (1994) présentée dans les travaux de Wiemer-Hastings et Zipitria (2001). Mais SELSA étend ce travail vers un cadre plus général où un mot avec un contexte syntaxique spécifié par ses mots adjacents est considéré comme une unité de représentation de connaissances. L'évaluation de cette approche a montré que la méthode LSA était plus pertinente que SELSA dans un test de corrélation avec des experts. Cependant, SELSA se révèle plus précise pour ce qui est de tester les bonnes et mauvaises réponses (c.-à-d. SELSA fait moins de fautes que LSA mais en retourne de plus nuisibles).

L'approche *ExpLSA* que nous présentons dans cet article se place dans un contexte différent. En effet, dans notre cadre de travail, les contextes sont représentés par des phrases. Ceux-ci ont donc une taille réduite ce qui a tendance à donner des résultats décevants avec l'utilisation de LSA (Rehder et al. (1998); Roche et Chauché (2006)). Dans notre approche, nous proposons d'utiliser la régularité de certaines relations syntaxiques afin d'enrichir le contexte comme nous allons le monter dans la section suivante.

4 Notre Approche : *ExpLSA*

Le but final que nous nous fixons consiste à regrouper automatiquement des termes extraits grâce à des systèmes tels que ACABIT (Daille (1994)), LEXTER (Bourigault (1993)), SYNTEX (Bourigault et Fabre (2000)), EXIT (Roche et al. (2004)). Dans notre cas, nous nous proposons de regrouper les termes nominaux extraits avec EXIT. Les termes extraits avec ce système sont des groupes de mots respectant des patrons syntaxiques (nom-préposition-nom, adjectif-nom, nom-adjectif, etc.). Par ailleurs, EXIT s'appuie sur une méthode statistique afin de classer les termes extraits et utilise une approche itérative pour construire des termes complexes.

La première étape de ce regroupement pour finalement construire une classification conceptuelle sera effectuée par *ExpLSA* dont le principe est décrit dans la section suivante.

4.1 Principe général d'*ExpLSA*

Notre approche vise à enrichir le corpus initial lemmatisé en faisant une expansion des phrases (d'où le nom *ExpLSA*) fondée sur une méthode syntaxique. Il en ressort un contexte plus riche. Celui-ci est construit en complétant les mots du corpus par des mots jugés sémantiquement proches.

Citons par exemple la phrase : *"Vos interlocuteurs seront donc bien inspirés de placer les échanges ..."*. Nous la transformons tout d'abord en phrase lemmatisée via le système SYGFRAN (Chauché (1984)) : *"Votre interlocuteur être donc bien inspiré de placer le échange*

..."*.* Enfin, elle va être enrichie avec d'autres mots en devenant la phrase *"Votre (interlocuteur collaborateur) être donc bien inspiré de placer le échange ..."*. Les méthodes utilisées pour déterminer les mots utilisés pour cet enrichissement, ainsi que la sélection de ceux-ci, sont présentées dans la section suivante.

4.2 L'analyse syntaxique pour mesurer la proximité sémantique

Afin d'enrichir le corpus initial, nous proposons d'effectuer tout d'abord une analyse syntaxique du corpus avec SYGFRAN. Ce dernier nous renvoie les relations syntaxiques présentes dans chaque phrase. Dans notre approche, nous nous sommes plus particulièrement intéressés aux relations syntaxiques Verbe-Objet (Verbe_Preposition_Complément, Verbe_COD) du corpus. Par exemple, nous avons extrait la relation syntaxique *"Verbe :intéresser, Objet :interlocuteur"* à partir de la phrase *"Ils intéressent leurs interlocuteurs."*.

Une fois l'ensemble des relations Verbe-Objet extraites, nous utilisons une mesure pour évaluer la proximité sémantique des verbes, la mesure d'Asium (Faure (2000); Faure et Nedellec (1999)). Cette mesure considère des verbes comme proches s'ils possèdent un nombre important d'objets en commun. Le principe de cette approche est similaire à celle présentée par Bourigault (2002).

Soient p et q, deux verbes avec leurs objets respectifs $p_1, ..., p_n$ et $q_1, ..., q_n$.$NbOccCom_p(q_i)$ représente le nombre d'occurrences des objets q_i en relation avec le verbe q qui sont aussi des objets du verbe p, $NbOcc(q_i)$ représente le nombre d'occurrences des objets q_i. La mesure d'Asium est définie de la manière suivante :

$$Asium(p,q) = \frac{log_{Asium}(\sum NbOccCom_q(p_i)) + log_{Asium}(\sum NbOccCom_p(q_i))}{log_{Asium}(\sum NbOcc(p_i)) + log_{Asium}(\sum NbOcc(q_i))}$$

Avec $log_{Asium}(x)$ valant :
- pour $x = 0$, $log_{Asium}(x) = 0$
- sinon $log_{Asium}(x) = log(x) + 1$

Une mesure d'Asium proche de 1 implique une importante proximité sémantique. L'exemple de la figure 1 illustre l'application de la mesure d'Asium pour les verbes *écouter* et *convaincre*. Nous considérerons par la suite plusieurs seuils de similarité, appelés SA, signifiant qu'au delà de ceux-ci, les verbes seront considérés comme proches par la mesure d'Asium. La méthode d'expansion utilisant la méthode d'Asium est décrite dans la section suivante.

FIG. 1 – *Mesure d'Asium entre les verbes écouter et convaincre*

4.3 Étapes d'*ExpLSA*

Après avoir explicité la mesure d'Asium permettant de mesurer la proximité des verbes du corpus, nous proposons de détailler les différentes étapes définissant *ExpLSA* afin d'étendre les contextes.

La première étape de l'approche *ExpLSA* identifie les différents termes extraits par EXIT. Cette identification consiste à représenter le terme par un seul mot (par exemple, le terme *attitude profondément participative* issu du corpus des Ressources Humaines devient *nom234* qui représente le 234ème terme parmi une liste extraite par EXIT).

Après extraction des relations syntaxiques Verbe-Objet par le biais d'une analyse syntaxique, la phase suivante de notre approche consiste à étudier la proximité sémantique entre les verbes en utilisant la mesure d'Asium (section 4.2). Chaque verbe du corpus est donc évalué avec tous les autres, en ne conservant que le couple ayant obtenu le meilleur score de similarité. Nous pourrons par exemple déduire que les verbes *écouter* et *convaincre* sont sémantiquement proches au sens d'Asium car ils partagent communément les objets *interlocuteur* et *collaborateur* (voir figure 1).

L'étape suivante a pour but de regrouper tous les objets communs dont les verbes ont été jugés proches sémantiquement par le seuil de similarité le plus élevé parmi l'ensemble des couples de verbes.

Nous considérons deux possibilités de regroupement. La première consiste à considérer les objets communs aux deux verbes (*interlocuteur* et *collaborateur* dans l'exemple de la figure 1). La seconde considère les objets communs et les complémentaires aux deux verbes comme dans les travaux de Faure et Nedellec (1999) (*interlocuteur, collaborateur, autrui, personne* dans l'exemple de la figure 1).

Nous proposons donc de compléter le corpus initial en attachant à chaque mot les autres mots communs. Ainsi, notre phrase initiale :

– *Votre **interlocuteur** être donc bien inspiré de placer le échange ...*

va devenir avec la première méthode d'expansion (méthode dite des intersections) :

– *Votre **(interlocuteur collaborateur)** être donc bien inspiré de placer le échange ...*

et va devenir avec la seconde méthode d'expansion (méthode dite des complémentaires) :

– *Votre **(interlocuteur collaborateur autrui personne)** être donc bien inspiré de placer le échange ...*

L'approche LSA peut alors être appliquée à partir de ce corpus enrichi.

Notons qu'une liste de noms non porteurs de sens ne sont pas pris en compte pour enrichir le contexte (par exemple, les mots "chose", "personne", etc.). Cette liste a été constituée manuellement.

L'évaluation, qui va être présentée dans la section suivante, consiste à comparer les résultats obtenus automatiquement en nous appuyant sur *ExpLSA* avec ceux d'un expert qui a associé manuellement les termes pertinents à des concepts.

5 Expérimentations

Pour discuter de la qualité des résultats retournés avec notre approche, nous nous appuyons sur le protocole expérimental décrit dans la section suivante.

5.1 Protocole expérimental

Dans nos expérimentations nous nous appuyons sur le corpus des Ressources Humaines expertisé manuellement. De cette expertise ressort une classification conceptuelle de l'ensemble des termes extraits par EXIT ; les concepts ayant été définis par l'expert. Par exemple, l'expert a défini le concept "Relationnel" dont les termes *confrontation ouverte*, *contact superficiel* et

entourage compréhensif sont des instances. L'objectif de nos expérimentations est d'évaluer les similarités entre les termes retournées par les méthodes automatiques ci-dessous :

- *M1* : LSA[4]
- *M2* : la méthode des intersections de *ExpLSA*
- *M3* : la méthode des complémentaires de *ExpLSA*
- *M4* : LSA + Tree-Tagger

La méthode de LSA + Tree-Tagger consiste à utiliser un étiqueteur grammatical, le Tree-Tagger (Schmid (1995)) comme dans l'approche de Wiemer-Hastings et Zipitria (2001) présentée dans la section 3. Ainsi, nous appliquons LSA sur un corpus qui a été au préalable étiqueté par le Tree-Tagger.

Pour comparer ces méthodes, nous avons évalué deux à deux les termes des concepts. Pour cela, nous avons sélectionné parmi les concepts, ceux étant les plus représentés dans le corpus, c'est-à-dire les concepts regroupant un minimum de 200 termes distincts selon l'expertise manuelle. Cela nous laisse un total de quatre concepts. Nous proposons ainsi de comparer deux à deux chaque concept. L'objectif de nos expérimentations consiste à évaluer si les termes appartenant à un concept sont correctement associés aux termes de ce même concept par les quatre méthodes (M1 à M4) décrites ci-dessus.

Pour effectuer une telle comparaison, tous les termes d'un concept C1 sont pris en compte. La similarité (cosinus) est alors calculée entre l'ensemble des termes du concept C1 et les autres termes des concepts à comparer (par exemple, les termes de C1 + C2, C1 + C3 ou C1 + C4). Les couples de termes ainsi obtenus sont classés par valeur décroissante avec les quatre méthodes M1 à M4. Un système retourne des résultats de bonne qualité si les couples pertinents sont placés en début de liste. Un couple est pertinent si les deux termes appartiennent au même concept. Pour évaluer la qualité de la liste, nous calculons la précision des premiers couples retrournés, le rappel n'ayant pas été jugé adapté[5]. La précision permet d'évaluer la proportion de couples pertinents retrouvés par le système. Notons que la réalisation de ces expérimentations avec deux concepts est assez conséquente puisqu'elle produit plus de $60\,000$ calculs de similarité.

5.2 Comparaison des deux méthodes *ExpLSA*

Cette première évaluation propose de comparer les deux méthodes de *ExpLSA* utilisées pour l'enrichissement du corpus, la méthode des intersections (M2) et la méthode des complémentaires (M3). Le tableau 1 compare la moyenne des précisions des 100[6] premiers couples

TAB. 1 – *Précision en fonction des 100 premiers couples de termes. M1 : LSA, M2 : ExpLSA avec la méthode des intersections, M3 : ExpLSA avec la méthode des complémentaires.* SA *= 0,6.*

[4]Nous utilisons le logiciel Infomap pour mener nos expérimentations, http ://infomap-nlp.sourceforge.net/

[5]Le but de notre approche est d'avoir les couples pertinents placés en début de liste. Cependant, dans le cas où peu de couples sont pris en compte ($n < 100$), le rappel est naturellement très faible et n'est pas nécessairement adapté pour évaluer la performance de notre approche.

[6]Cette valeur de 100 a été établie car elle correspond à un nombre raisonnable de couples proposés à un expert.

de termes avec un seuil pour la mesure d'Asium à 0,6. Avec une telle valeur de SA, nous faisons une large expansion du corpus. Ce tableau montre que les approches *ExpLSA* utilisant la méthode des intersections (M2) ou bien celle des complémentaires (M3) sont inférieures à LSA et n'améliorent que rarement la précision. Ces résultats s'expliquent par la quantité importante de données non pertinentes utilisées pour l'enrichissement. En effet, un seuil SA faible a tendance à ajouter du bruit comparativement à l'utilisation d'un seuil plus élevé qui permet une expansion quantitativement plus faible mais souvent plus pertinente. Ceci confirme les résultats préliminaires présentés dans (Béchet et al. (2007)). Le meilleur compromis entre la quantité de données ajoutées et la qualité de celles-ci a été experimentalement établi avec un seuil SA égal à 0,8 sur notre corpus. Nous constatons par ailleurs que la méthode M3 donne globalement des résultats plus faibles que la méthode M2 (bien que toutes les deux soient inférieures à M1).

La table 2 montre les mêmes expérimentations que pour la table 1 en utilisant SA à 0,8. Ce

TAB. 2 – *Précision en fonction des 100 premiers couples de termes. M1 : LSA, M2 : ExpLSA avec la méthode des intersections, M3 : ExpLSA avec la méthode des complémentaires.* SA *= 0,8.*

tableau montre de meilleurs résultats pour la méthode M2 et améliore les résultats de LSA dans trois cas sur six. Les trois cas où LSA obtient la meilleure précision incluent le concept *Relationnel*. *Relationnel* peut-être sémantiquement proche des autres concepts avec une frontière plus difficile à identifier de manière automatique (par exemple, les concepts *Relationnel* et *Comportement / Attitude*). Ainsi, sans considérer ce concept qui peut se réveler dans certains cas ambigu, notre méthode améliore LSA. Des expérimentations avec d'autres corpus exploitant des concepts plus discriminants devront être menées pour confirmer ces résultats.

Enfin la figure 2 confirme le fait que la méthode des complémentaires (M3) donne des résultats de moins bonne qualité par rapport à celle des intersections (M2). Nous conserverons donc uniquement la méthode des intersections dans les prochaines expérimentations.

FIG. 2 – *Précision en fonction des 100 premiers termes comparant les deux méthodes M2 et M3 propres à* ExpLSA *avec les concepts Comportement / Attitude et Environnement*

5.3 *ExpLSA comparé à la méthode LSA + Tree-Tagger*

TAB. 3 – *Précision en fonction des 100 premiers couples de termes. M1 : LSA, M2 : ExpLSA (méthode des intersections et* SA *= 0,8) , M4 : LSA + Tree-Tagger.*

Nous proposons de comparer dans cette section deux méthodes utilisant des connaissances syntaxiques, *ExpLSA* (M2) et LSA + Tree-Tagger (M4). LSA + Tree-Tagger consiste à ajouter

des connaissance grammaticales au corpus en complétant les mots par une étiquette grammaticale. Cette approche permet de lever les ambiguités de certains mots pouvant appartenir à des catégories grammaticales différentes. Par exemple, le mot *bien* peut-être un adverbe, un nom ou un adjectif. Avec la méthode M4, nous considérons dans cet exemple trois formes distinctes pour représenter ce mot : *bien_ADV*, *bien_NOM* et *bien_ADJ*.

La figure 3 montre que la méthode LSA + Tree-Tagger améliore les résultats de LSA pour les derniers couples ce qui ne correspond pas aux attentes de l'utilisateur. En effet, une fonction de rang est en général satisfaisante si un nombre important d'exemples positifs sont placés en tête de liste. Cette tendance se généralise pour les autres concepts comme le montre le tableau 3. La méthode LSA + Tree-Tagger (M4) reste cependant presque toujours inférieure à notre approche *ExpLSA* (M2). Ces résultats nous encouragent à envisager dans de futurs travaux une hybridation des méthodes M2 et M4 afin de conserver les résultats de *ExpLSA* et de bénéficier des améliorations de la méthode LSA + Tree-Tagger pour les derniers couples.

FIG. 3 – *Précision en fonction des 100 premiers couples de termes comparant les méthode ExpLSA, LSA et LSA + Tree-Tagger pour les concepts Comportement / Attitude et environnement*

6 Conclusion et discussion

LSA est une méthode statistique utilisée notamment pour regrouper des termes afin d'établir une classification conceptuelle. Néanmoins, cette méthode donne des résultats parfois décevants. Ceux-ci s'expliquent entre autres par l'absence de connaissances linguistiques. La qualité de ces résultats peut également être influencée par la taille des contextes utilisés, LSA obtenant de meilleurs résultats avec des contextes de grande taille.

C'est pourquoi nous nous sommes intéressés dans nos travaux à améliorer les performances de LSA avec des contextes assez courts (phrases) en proposant une approche, *ExpLSA*, consistant à effectuer une expansion des contextes avant d'appliquer LSA. Nous rendons de ce fait les contextes plus riches en utilisant des outils syntaxiques afin d'y parvenir.

Nous avons présenté deux expérimentations pour comparer l'approche *ExpLSA*. Nous avons conclu dans la première expérience qu'avec *ExpLSA* fondée sur la méthode des complémentaires, l'expansion réalisée n'était pas pertinente et ajoutait une quantité importante de bruit. *ExpLSA* utilisant la méthode des intersections donne quant à elle des résultats satisfaisants pour les concepts discriminants d'un point de vue sémantique. Les termes des concepts pouvant générer des ambiguités se révelent plus ou moins difficiles à traiter automatiquement par notre approche. L'inconvénient de la méthode $ExpLSA$ est qu'elle nécessite un temps d'exécution conséquent (environ deux heures pour traiter un corpus de 1,2 Mo). Ce temps important s'explique principalement par la durée d'execution de la tâche d'extraction des relations syntaxiques. La seconde expérimentation a montré que LSA + Tree-Tagger améliore rarement LSA et que notre approche *ExpLSA* donne de meilleurs résultats.

Nous envisageons comme futurs travaux d'approfondir les expérimentations en identifiant plus précisément dans quels cas $ExpLSA$ donne de meilleurs résultats comparativement à LSA. Ceci permettra de mettre en place une approche hybride qui utilise LSA et/ou $ExpLSA$

et/ou LSA + Tree-Tagger selon les situations les plus appropriées. De plus, nous souhaiterions valider le regroupement des mots avec $ExpLSA$ en nous appuyant sur des mesures statistiques et des données numériques issues des moteurs de recherche du web (Turney (2001)). Par ailleurs, nous proposerons d'autres méthodes afin d'ajouter des connaissances syntaxiques à LSA. De plus, nous validerons notre méthode d'expansion des contextes en la confrontant à des problèmes de classification de textes. Nous envisageons enfin d'utiliser des vecteurs sémantiques avec SYGMART (Chauché (1984)) en considérant un terme comme produit d'un ensemble de concepts issus du thésaurus Larousse.

Remerciements

Nous remercions Yves Kodratoff (Equipe IA, LRI, France) et Serge Baquedano (société PerformanSe) pour leur travail d'expertise réalisé sur le corpus des ressources humaines.

Références

Béchet, N., M. Roche, et J. Chauché (2007). Improving LSA by expanding the contexts. In *Context-Based Information Retrieval (CIR) workshop - CONTEXT'07*, pp. 105–108.

Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *T.A.L. 34*(2), 105–118.

Bourigault, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN, Nancy*, pp. 75–84.

Bourigault, D. et C. Fabre (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires 25*, 131–151.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, pp. 722–727.

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

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

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph. D. thesis, Université Paris 7.

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. Nedellec (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : The system ASIUM. In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management, number 1937 in LNAI*, pp. 329–334.

Kanejiya, D., A. Kumar, et S. Prasad (2003). Automatic evaluation of students' answers using syntactically enhanced lsa. In *Proceedings of the Human Language Technology Conference (HLT-NAACL 2003) Workshop on Building Educational Applications using NLP*.

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., D. Laham, B. Rehder, et M. E. Schreiner (1997). How well can passage meaning be derived without using word order ? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pp. 412–417.

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. et J. Chauché (2006). LSA : les limites d'une approche statistique. In *Actes de l'atelier FDC'06 (Fouille de Données Complexes), conférence EGC'2006*, pp. 95–106.

Roche, M., T. Heitz, O. Matte-Tailliez, et Y. Kodratoff (2004). EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04*, Volume 2, pp. 946–956.

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)*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop, Dublin*.

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. et I. Zipitria (2001). Rules for syntax, vectors for semantics. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*.

Summary

Latent Semantic Analysis (LSA) is nowadays used in various fields like cognitive models, educational applications but also in classification. We propose in this paper an approach which adds grammatical knowledge to LSA. Different methods are studied to finally perform a conceptual classification.