Analyse sémantique latente et segmentation automatique des textes

Yves Bestgen

FNRS – UCL/PSOR –1348 Louvain-la-Neuve – Belgique
yves.bestgen@psp.ucl.ac.be

Abstract

Latent semantic analysis (LSA) is employed in psycholinguistics and in computational linguistics to acquire domain-independent semantic knowledge which is then used to model cognitive processes or to develop automated text analysis technologies. By analyzing the effectiveness of an algorithm of text segmentation which rests on LSA, we show not only that a generic semantic space is less effective than a specific space, but also that a series of parameters, which must be fixed at the time of the acquisition phase of the semantic knowledge, affect the performance of the algorithm.

Résumé

L'analyse sémantique latente (ASL) est employée tant en psycholinguistique qu'en linguistique computationnelle pour acquérir des connaissances sémantiques indépendantes de tout domaine qui sont ensuite employées pour modéliser des processus cognitifs ou développer des techniques d'analyse automatique du langage. En analysant l'efficacité d'un algorithme de segmentation des textes qui repose sur l'ASL, nous montrons non seulement qu'un espace sémantique générique est moins efficace qu'un espace spécifique, mais aussi qu'une série de paramètres qui doivent être fixés lors de la phase d'acquisition des connaissances sémantiques affectent les performances de l'algorithme[1].

Mots-clés : segmentation automatique, analyse sémantique latente, connaissances indépendantes de tout domaine, lemmatisation.

1. Introduction

Depuis quelques années, l'analyse sémantique latente (ASL ou LSA : Latent Semantic Analysis) est à la base d'un nombre de plus en plus important de recherches en psycholinguistique (Landauer *et al.*, 1998). Cette technique vise à construire un espace sémantique de très grande dimension à partir de l'analyse statistique de l'ensemble des cooccurrences dans un corpus de textes. Son succès en psychologie s'explique par son utilisation pour développer des simulations des processus psycholinguistiques à l'œuvre lors de la compréhension du langage (Landauer et Dumais, 1997), incluant, par exemple, un « modèle computationnel » du traitement des métaphores (Kintsch, 2000 ; Lemaire *et al.*, 2001), mais aussi l'analyse de la cohérence dans des textes (Bestgen *et al.*, 2003 ; Foltz *et al.*, 1998).

Cette technique n'est pas, tant s'en faut, l'apanage de la psychologie. Elle repose en effet sur la décomposition en valeurs singulières, une propriété des matrices rectangulaires, proposée par Eckart et Young dès 1936, qui est à la base des méthodes factorielles d'analysés de

[1] Cette recherche a bénéficié du soutient de la Communauté française de Belgique – Actions de recherche concertées et du Fonds de la Recherche fondamendale collective (FRFC/FNRS).

données (Lebart, Morineau et Piron, 2000). Plus récemment, la psychologie a emprunté cette technique à l'analyse automatique du langage (Deerwester *et al.*, 1990). Dans ce champ de recherche, l'intérêt principal de l'analyse sémantique latente est de permettre la construction automatique de connaissances sémantiques génériques (*domain-independent knowledge*). Ces connaissances peuvent donc être employées quel que soit le domaine dont sont issus les textes pour développer des techniques d'analyse du langage comme celles employées en indexation automatique de documents, en recherche de l'antécédent d'une anaphore, en identification automatique d'expressions idiomatiques ou encore en segmentation des textes (Choi *et al.*, 2001 ; Degand et Bestgen, 2003 ; Klebanov et Wiemer-Hastings, 2002). En psychologie également, la thèse selon laquelle il est possible de constituer un espace sémantique générique apte à modéliser de nombreux processus mentaux est aussi fréquemment mise en avant (Kinstch, 1998 ; Landauer et Dumais, 1997). A notre connaissance, cette thèse n'a jamais fait l'objet d'une vérification empirique. Bien plus, plusieurs recherches ont été menées en recourant à des bases spécifiques au domaine dont était issu le matériel (par exemple Bestgen et Cabiaux, 2002 ; Wolfe *et al.*, 1998). Etudier la validité de cette thèse est un des objectifs principaux de la présente étude. Nous souhaitons toutefois aller au-delà de cette seule question en étudiant l'impact d'autres paramètres qui doivent être fixés lors de la construction d'un espace sémantique et qui sont pourtant rarement discutés par les auteurs.

Pour aborder ces questions, le champ de la segmentation automatique des textes nous semble particulièrement propice pour les raisons suivantes. Tout d'abord ; l'opposition entre connaissances spécifiques à un domaine et connaissances génériques y est particulièrement prégnante (Ferret, 2002). De plus, la segmentation automatique des textes est un domaine de recherche en plein développement (Manning et Schütze, 1999 : 572) et, récemment, Choi *et al.* (2001) ont montré que l'ASL permettait le développement d'un algorithme de segmentation automatique plus efficace que plusieurs procédures classiques. Notons toutefois que ces auteurs n'ont eu recours qu'à un espace sémantique spécifique au matériel employé pour tester leur algorithme. L'analyse de la cohérence/segmentation des textes est tout aussi importante pour la psycholinguistique puisqu'identifier les relations de cohérences, mais aussi les ruptures thématiques dans un texte, est une composante centrale de la compréhension (Bestgen et Vonk, 2000 ; Gernsbacher, 1990 ; Kintsh, 1998). Enfin, les méthodes de segmentation automatique fournissent un critère relativement objectif pour jauger l'efficacité d'un espace sémantique donné, critère beaucoup plus difficile à définir dans les études psycholinguistiques.

La suite de ce rapport est structurée de la manière suivante. Après avoir décrit l'analyse sémantique latente et les paramètres qui doivent être fixés lors de la construction d'un espace sémantique, nous présentons l'algorithme de segmentation de Choi. Ensuite, nous rapportons une expérience visant à déterminer l'impact d'un ensemble de paramètres sur l'efficacité de la segmentation. Dans la conclusion, les limites de la présente étude et des pistes dé développement sont discutées.

2. L'analyse sémantique latente

Le point de départ d'une analyse sémantique latente consiste en un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chacun des documents, un document pouvant être un texte, un paragraphe ou même une phrase. Pour dériver d'un tableau lexical les relations sémantiques entre les mots, la simple analyse des cooccurrences brutes se heurte à un problème majeur. Même dans un grand corpus de textes, la plus grande partie des mots sont relativement rares. Il s'ensuit que les cooccurrences le sont encore plus. Leur rareté les rend

particulièrement sensibles à des variations aléatoires (Burgess *et al.*, 1998 ; Rajman *et al.*, 1997). L'ASL résout ce problème en remplaçant le tableau de fréquences original par une approximation qui produit une sorte de lissage des associations. Pour cela, le tableau de fréquences fait l'objet d'une décomposition en valeurs singulières avant d'être recomposé à partir d'une fraction seulement de l'information qu'il contient. Les milliers de mots caractérisant les documents sont ainsi remplacés par des combinaisons linéaires ou 'dimensions sémantiques' sur lesquelles peuvent être situés les mots originaux. Contrairement à une analyse factorielle classique, les dimensions extraites sont très nombreuses (plusieurs centaines) et non interprétables. Elles peuvent toutefois être vues comme analogues aux traits sémantiques fréquemment postulés pour décrire le sens des mots (Landauer *et al.*, 1998). Les différentes étapes nécessaires pour dériver un espace sémantique d'un tableau lexical sont illustrées dans l'Annexe 1.

Tant les mots que les segments originaux sont positionnés dans cet espace sémantique, ce qui permet de mesurer leur proximité. Plus précisément, le sens de chaque mot y est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, on calcule le cosinus entre les vecteurs qui les représentent. Les mêmes calculs peuvent être effectués sur les vecteurs qui représentent les documents analysés. Le plus important toutefois est que cette technique est encore plus générale puisqu'elle permet de calculer le vecteur qui correspond à un groupe de mots même si ce groupe de mots ne constitue pas un document analysé en tant que tel. Il est ainsi possible d'analyser la proximité sémantique entre deux phrases que celles-ci fassent partie du corpus de départ ou non, que le corpus de départ ait été segmenté en documents correspondant à des phrases ou non.

Une série de décisions, rarement discutées dans les travaux basés sur l'ASL (mais voir Lebart et Salem (1994) pour une discussion dans le domaine de la statistique textuelle), doivent être prises dès la constitution du tableau lexical. La première porte sur la nature des documents analysés : textes entiers comme des articles de journaux ou d'encyclopédies, paragraphes, phrases, unités arbitraires de tailles constantes. A cela s'ajoute la décision de lemmatiser ou non le corpus, un facteur connu pour affecter la classification automatique des textes (Riloff, 1995), d'éliminer ou non les mots très fréquents et les mots très rares (Dumais, 1995). Selon nous, l'explication la plus probable de l'absence de discussion de ces paramètres réside dans la croyance érigée en postulat que le très grand nombre de textes et de mots qui interviennent dans ce genre d'analyse rend peu probable l'existence d'effets liés à ces paramètres. Pour vérifier le bien-fondé de cette assertion et évaluer l'impact du caractère spécifique ou non de l'espace sémantique, nous avons réalisé une expérience dans laquelle des espaces sémantiques construits en faisant varier ces paramètres ont été employés comme source de connaissance pour segmenter des textes au moyen de la méthode proposée par Choi *et al.* (2001).

3. La segmentation des textes et l'algorithme de Choi

Depuis une dizaine d'années, de nombreuses méthodes ont été proposées pour segmenter automatiquement des textes. Elles se distinguent principalement par le type d'indices employés. Certaines se basent exclusivement sur une analyse de la cohésion lexicale alors que d'autres prennent également en compte des dispositifs linguistiques qui ont pour fonction de signaler la présence de changement de thèmes. Une autre distinction importante oppose les approches qui s'appuient exclusivement sur les informations contenues dans le texte à segmenter et celles qui ont recours à des connaissances acquises par ailleurs. La méthode proposée par Choi (2000 ; Choi *et al.*, 2001) se base exclusivement sur la cohésion lexicale, mais existe dans deux versions correspondant à ce second critère de différenciation.

La procédure de Choi est composée de trois étapes. Tout d'abord, le document à segmenter est découpé en unités textuelles minimales, habituellement les phrases. Les mots composant ces phrases sont soumis à différents traitements comme la suppression de mots peu informatifs sur le thème du texte (article, pronom, verbes très fréquents, ...) et une lemmatisation. Ensuite, une mesure de similarité entre toutes les paires d'unités prises deux à deux est calculée. Enfin, le document est segmenté de façon récursive en fonction des frontières entre les unités textuelles qui maximisent la somme des similarités moyennes à l'intérieur des segments ainsi constitués.

L'étape la plus importante pour la présente étude est celle qui calcule la similarité entre toutes les paires de phrases prises deux à deux. La procédure initialement proposée par Choi (2000) reposait sur la métrique du cosinus appliquée aux vecteurs représentant les paires de phrases (Manning et Schütze, 1999 : 539 et suiv.) . Pour être déclarés cohérents, deux passages doivent contenir des mots communs. Il s'agit d'une conception très restrictive de la cohésion lexicale. Afin de dépasser cette limitation, Choi *et al.* (2001) ont proposé d'employer l'analyse sémantique latente pour estimer la similarité entre deux phrases. Pour ce faire, on applique la métrique du cosinus non aux vecteurs bruts, mais aux vecteurs pondérés par les dimensions sémantiques dérivées par l'analyse sémantique latente (Manning et Schütze, 1999 : 554 et suiv.). Les étapes ultérieures sont identiques quelle que soit la méthode employée pour calculer les similarités.

Dans une première évaluation réalisée sur un matériel très spécifique (voir ci-dessous), Choi (2000) et Choi *et al.* (2001) ont montré que leurs méthodes étaient plus efficaces que plusieurs autres approches telles que *TextTilling* de Hearst, *DotPlot* de Reynar, *Segmenter* de Kan, Klavans et McKeown et le *Maximum-probability segmentation algorithm* de Utiyama et Isahara.

4. Expérience

L'ensemble des analyses rapportées ici a été effectué sur la base d'un corpus d'articles parus dans le journal belge francophone *Le Soir* en 1997 (plus de 24 000 000 de mots). Le corpus de départ a été divisé en deux périodes : de janvier à juin (S1) et de juillet à décembre (S2). Comme les articles étaient de longueurs très variables, nous avons éliminé de chaque période les articles faisant partie des 10% les plus courts ou des 10% les plus longs, soit ceux de moins de 36 et de plus de 2219 mots dans S1 et de moins de 41 et de plus de 2252 dans S2.

4.1. *Matériel pour les tests*

Pour déterminer l'impact des paramètres décrits ci-dessus sur la précision de la segmentation, nous avons employé la méthodologie, devenue classique, avec laquelle Choi a évalué ses algorithmes (Ferret, 2002 ; Utiyama et Isahara, 2001). Elle consiste à retrouver les frontières entre des textes qui ont été concaténés. Chaque ensemble de tests est composé de 100 échantillons. Chaque échantillon est composé de 10 segments de textes. Chaque segment est composé des *n* premières phrases d'un texte sélectionné aléatoirement dans le corpus de départ. Pour tester ses algorithmes, Choi a fait varier le paramètre *n* en prenant les 3 à 5, les 6 à 8 ou les 9 à 11 premières phrases. Comme les résultats furent quasiment identiques quelle que fût la valeur du paramètre *n*, nous n'avons employé qu'un seul des trois couples : les 9 à 11 premières phrases. La valeur du paramètre *n* pour chaque segment de texte est déterminée aléatoirement. Pour chaque demi-année du Soir, deux ensembles de test, composés chacun de 100 échantillons, ont été constitués par une procédure aléatoire entièrement automatisée.

4.2. *Manipulation des paramètres pour la constitution des espaces sémantiques*

1) *Base spécifique ou non spécifique*. Les bases sémantiques étaient constituées soit à partir de S1, soit à partir de S2. Lorsque le matériel de test a été extrait des articles qui ont servi à construire la base sémantique, on parlera de condition spécifique. Dans le cas inverse, par exemple matériel de test extrait des 6 derniers mois et base construite à partir des 6 premiers mois, on parlera de condition non spécifique.
2) *Unité textuelle pour la constitution de la base*. Trois types de découpages ont été testés : textes (Texte), paragraphes[2] (Parag.) ou unités arbitraires de taille constante. Dans ce dernier cas, trois tailles arbitraires ont été analysées : unité de 50 mots (Arb. 50), unité de 250 mots (Arb. 250), unité de 375 mots (Arb. 375). La première correspond approximativement à la taille moyenne d'un paragraphe et la dernière à celle d'un texte. Les unités arbitraires respectaient les limites des textes.
3) *Lemmatisation ou non.* Nous avons employé le programme « TreeTagger » de Schmid (1994) pour lemmatiser les corpus.
4) *Suppression des mots très rares*. Trois niveaux ont été définis pour ce paramètre : absence de suppressions[3], suppression des mots dont la fréquence totale dans le corpus est inférieure à 10 ou à 20.

Un cinquième paramètre aurait pu être manipulé : la suppression ou non des mots fréquents. Il ne l'a pas été parce que les procédures de segmentation automatique des textes prévoient systématiquement la suppression de mots peu informatifs parce que très fréquents ou appartenant à des classes fermées (article, pronom, ...). Nous avons donc choisi de toujours supprimer ces mots tant dans le matériel-test que lors de la construction des espaces sémantiques. Nous n'avons pas non plus manipulé le nombre de vecteurs conservés dans l'espace sémantique réduit, le fixant à 300 une valeur classique pour ce genre d'analyse (Landauer *et al.*, 1998).

4.3. *Implémentation*

Les décompositions en valeurs singulières ont été effectuées au moyen du programme SVDPACKC (Berry, 1992 ; Berry *et al.*, 1993). L'algorithme C99 de Choi (2000) et la variante incluant l'analyse sémantique latente (Choi *et al.*, 2001) ont été implémentés en C. Il n'a pas été possible de valider spécifiquement la version ASL de l'algorithme parce que nous ne disposons pas de l'espace sémantique employé par Choi *et al.* (2001). Par contre, la version C99, qui emploie les vecteurs de mots bruts, a pu être validée en comparant les résultats obtenus avec notre implémentation à ceux rapportés par Choi (2000) pour une partie de son matériel. Une correspondance parfaite a été obtenue.

5. Analyses et résultats

Afin de simplifier les analyses, nous avons employé la variante de l'algorithme de Choi dans laquelle le nombre de segments à trouver est imposé. Pour évaluer l'exactitude d'une segmentation, nous avons utilisé les indices classiques de précision et de rappel qui, dans le cas présent, sont égaux puisqu'on impose à l'algorithme de produire le nombre correct de segments[4].

[2] Nous avons éliminé les 10% des paragraphes les plus courts et les 10% les plus longs.

[3] Plus exactement, tous les mots dont la fréquence dans le corpus est égale ou supérieure à 2 sont pris en compte, les hapax ne pouvant pas être analysés.

[4] Les analyses ont également été effectuées sur la métrique d'erreur proposée par Beefermn *et al.* (1998) et

Les analyses ont été effectuées en deux temps. Tout d'abord, nous avons étudié l'effet de la taille des documents sur l'espace sémantique en maintenant constants tous les autres paramètres sauf le caractère spécifique ou non de la base. Cette première analyse a montré que le découpage en textes donnait le meilleur résultat. Dés lors, nous avons choisi de n'effectuer la deuxième analyse que sur les découpages en textes. Lors de celle-ci, nous avons comparé dans un design factoriel complet les paramètres *Spécificité de la base*, *Lemmatisation* et *Suppression des mots rares*. La principale justification de cette procédure en deux temps réside dans le temps-calcul nécessaire pour constituer un espace sémantique.

5.1. *Analyse 1 : Impact de la taille des unités*

La première analyse porte sur l'effet de la nature et de la longueur des documents employés pour constituer l'espace sémantique. Cinq types de documents ont été comparés : subdivision en textes, en paragraphes et en unité de tailles constantes et arbitraires de 50 mots, de 250 mots et de 375 mots. Dans cette analyse, le caractère spécifique ou non de la base a également été pris en compte. Par contre, tous les autres paramètres ont été maintenus constants : absence de lemmatisation et seuil de fréquence pour les mots rares à 10. Les valeurs de précision pour les 400 échantillons de test ont été analysées au moyen d'une analyse de variance avec comme facteurs répétés la spécificité et le type de documents. Vu le grand nombre de données disponibles, un seuil de signification 10 fois plus strict que le classique 0.05 a été choisi pour toutes les analyses rapportées ici ($p<=0.005$).

	Documents				
	Texte	Parag.	Arb. 50	Arb. 250	Arb. 375
Spécifique	0.82	0.47	0.39	0.80	0.78
Non Spécifique	0.68	0.41	0.39	0.64	0.67

Tableau 1. Précision en fonction du type de document et de la spécificité de la base

On observe un effet très significatif du type d'unité et de la spécificité ainsi qu'une interaction entre ces deux facteurs. Comme le montre le tableau 1, les découpages en petites unités, qu'elles soient arbitraires ou non, donnent lieu à des performances très faibles. Il est à noter que Choi *et al.* (2001) ont obtenu, en anglais, des résultats nettement supérieurs avec des unités de cette taille. On observe peu de différences entre les trois autres unités, même si le découpage en texte donne lieu à des performances légèrement supérieures. L'impact de la spécificité de la base est également très important à l'avantage d'une base spécifique. L'interaction est largement provoquée par l'absence d'effet du facteur spécificité pour les unités arbitraires de 50 mots.

La conclusion principale de cette première analyse est qu'un découpage en grandes unités semble préférable. Il apparaît aussi que la variabilité de la taille des unités de type texte ne nuit pas à la performance. En effet, les meilleures performances sont obtenues avec ces unités dont la longueur varie fortement. Si des unités de ce type ne sont pas disponibles, par exemple lorsque le corpus est composé de textes très longs comme des romans uniquement découpés en chapitres, un découpage en unité arbitraire d'une taille suffisante produit des résultats satisfaisants. Notons enfin qu'en français aussi l'algorithme de Choi *et al.* (2001) atteint un niveau de performance très élevé.

employée par Choi, sans que les résultats ne soient modifiés d'une façon significative.

5.2. *Analyse 2*

Seul le découpage en texte a été employé pour cette analyse. Les 3 autres paramètres ont été croisés les uns avec les autres dans un design factoriel complet : *Spécificité de la base* (2) X *Lemmatisation* (2) X S*uppression des mots rares* (3), soit un total de 12 conditions correspondant à autant de bases sémantiques pour chaque période (S1 et S2). Ce même plan factoriel a été employé pour analyser les valeurs de précision obtenues pour les 400 échantillons de test au moyen d'une analyse de variance pour mesures répétées.

On obtient, comme lors de la première analyse un effet très important du paramètre *Spécificité de la base*. On observe aussi un effet significatif, mais nettement moins important, du seuil de suppression des mots rares.

Si la lemmatisation seule n'a pas d'effet, il existe une interaction significative entre ce facteur et la spécificité de la base. Comme le montre le tableau 2, la lemmatisation a un léger effet négatif lorsque les bases sont spécifiques (0.82 versus 0.80) alors qu'elle a un léger effet positif lorsqu'elles ne le sont pas (0.69 versus 0.68). Enfin, on observe une interaction significative entre la lemmatisation et la suppression des mots rares. On peut l'interpréter de la manière suivante : la suppression des mots rares a beaucoup moins d'effet lorsque le matériel a été lemmatisé. Ce résultat semble logique puisque l'effet principal de la lemmatisation est de réduire le nombre de formes rares en les regroupant sous les mêmes lemmes.

	Lemmatisation					
	Non			Oui		
	Seuil pour la suppression des mots rares					
	≥2	≥10	≥20	≥2	≥10	≥20
Spécifique	0.83	0.82	0.81	0.80	0.80	0.80
Non Spécifique	0.68	0.68	0.67	0.69	0.69	0.69

Tableau 2. Précision en fonction de la spécificité de la base, de la lemmatisation et du seuil pour la suppression des mots rares

5.3. *Analyse complémentaire*

Les deux résultats les plus importants de cette étude sont l'impact de la taille des unités et de la spécificité de la base. Si le premier s'interprète aisément, le second mérite une analyse complémentaire. En effet, deux explications différentes peuvent être données à cet effet. Rappelons d'abord qu'une même base est dans certaines analyses considérée comme spécifique et dans d'autres analyses considérée comme non spécifique selon l'ensemble de tests employé. Pour un ensemble de tests donné, les deux bases peuvent se différencier au niveau des dimensions sémantiques extraites, mais aussi au niveau des mots qu'elles contiennent. Il est donc possible que certains mots utiles pour la segmentation ne soient présents que dans la base spécifique. L'autre possibilité est que les relations sémantiques soient partiellement différentes dans les deux bases, que celles de la base spécifique soient plus adéquates pour traiter l'ensemble de test. Cette seconde explication nous semble beaucoup plus dommageable pour l'ASL puisqu'il ne suffirait pas d'augmenter le nombre de mots indexés dans une base pour la rendre efficace quelles que soient les analyses prévues.

Pour trancher entre ces alternatives, une analyse complémentaire a été effectuée. Son principe est simple : égaliser les mots employés pour mesurer la proximité entre les phrases. De cette

manière, seules les relations sémantiques dans les bases peuvent jouer. Nous avons donc effectué de nouvelles analyses dans lesquelles seuls les mots présents simultanément dans chaque paire[5] de bases spécifiques et non spécifiques sont utilisés pour le calcul des proximités. Les résultats indiquent nettement que ce ne sont pas les mots sélectionnés pour le calcul des proximités qui jouent, mais bien les relations sémantiques dans les bases. Lorsque les mêmes mots sont employés pour les espaces spécifiques et non spécifiques, la précision moyenne est de 0.81 dans la condition spécifique et de 0.68 dans la condition non spécifique, valeurs identiques à celles obtenues dans les analyses précédentes (voir tableau 2).

6. Conclusion

Avant de discuter d'une manière plus approfondie les résultats de cette étude, il est nécessaire de souligner plusieurs limites de la présente recherche qui mettent principalement en cause la généralité des résultats obtenus. Nous n'avons testé qu'une tâche (la segmentation), qu'une seule méthode (celle de Choi) dans un seul type de textes (des articles de journaux) et avec une seule méthodologie de test (la segmentation de fragments de documents arbitrairement concaténés). Même si la situation test résultant de toutes ces particularités correspond à celle employée par Choi (2000 et 2001) et par Utiyama et Isahara (2001), étudier l'impact des paramètres sur d'autres tâches et la segmentation avec d'autres types de textes et un autre matériel de test est nécessaire pour affermir l'argumentation.

Notre étude n'a pas non plus apporté de réponses à deux questions importantes. Tout d'abord, notre manipulation de la spécificité des bases sémantiques est très « légère » puisque nous avons opposé des bases construites à partir d'articles parus durant deux périodes successives d'un même journal. Cela a été suffisant pour produire des différences de performance très importantes. Jusqu'à quel niveau d'inefficacité serions-nous descendus si nous avions employé des journaux différents ou des genres de textes différents (un journal et une encyclopédie ou des œuvres littéraires) ? Des études complémentaires sont ici aussi nécessaires.

La deuxième question porte sur l'existence de différences importantes en fonction de la langue dans laquelle l'étude est menée. Comme indiqué plus haut, Choi *et al.* (2001) ont obtenu des résultats excellents avec des bases sémantiques dont les documents étaient des paragraphes. Dans notre étude, la construction d'un tableau lexical mots x paragraphes a produit des résultats médiocres. Nous avons également observé que l'algorithme C99 de Choi, qui ne s'appuie pas sur l'ASL et qui fonctionne très bien en anglais, produit en français des résultats très mauvais. Dans l'ensemble des essais effectués en variant les procédures de suppression des mots fréquents et de lemmatisation, la meilleure précision obtenue était inférieure à 0.60. Une différence d'efficacité entre l'anglais et le français a également été observée par Ferret (2002) lorsqu'il employa son système Topicoll pour segmenter des articles du Monde ou le matériel de Choi. Avant d'essayer d'expliquer ces observations par des différences inter-langues, plusieurs autres explications devraient être testées. Malgré les diverses tentatives effectuées, il est possible que les pré-traitements du matériel effectués par Choi soient plus efficaces que les nôtres. Or, ces pré-traitements affectent fortement les performances de l'algorithme tant en français qu'en anglais. L'effet inter-langues pourrait aussi résulter de différences au niveau du matériel à segmenter employé dans chacune des langues.

[5] Par paire de bases sémantiques, on entend ici deux bases identiques pour les paramètres lemmatisation et suppression des mots rares, mais extraites soit de S1, soit de S2.

En résumé, l'ensemble des paramètres manipulés dans cette étude a affecté l'efficacité de l'algorithme de segmentation. Les deux facteurs les plus importants sont le type de documents employé pour construire le tableau lexical et la spécificité de la base sémantique par rapport au matériel à segmenter. La lemmatisation et la suppression des mots rares ont aussi eu un effet, soit seul, soit en interaction. Il faut toutefois noter que l'effet de la lemmatisation est très faible, observation qui rejoint les conclusions de Lebart et Salem (1992, voir par exemple : 225-226). Dans le cas présent, elle réduit même l'efficacité de l'algorithme. Plus généralement, le fait d'analyser de très grands corpus de textes ne suffit pas à neutraliser les effets de ces paramètres.

Références

Beeferman D., Berger A. et Lafferty J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, vol. (34) : 177–210.

Berry M., Do T., O'Brien G., Krishna V. et Varadhan S. (1993). *SVDPACKC: Version* 1.0 *User's Guide*. Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

Berry M.W. (1992). Large scale singular value computation. *International journal of Supercomputer Application*, vol. (6) : 13-49.

Bestgen Y. et Cabiaux A.F. (2002). L'analyse sémantique latente et l'identification des métaphores. In *Actes de TALN 2002* : 331-337.

Bestgen Y., Degand L. et Spooren W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an explorative study. In Lagerwerf L., Spooren W. et Degand L. (Eds), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse 2003*. Nodus Publikationen : 189-202.

Bestgen Y. et Vonk W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language*, vol. (42) : 74-87.

Burgess C., Livesay K. et Lund K. (1998). Explorations in Context Space : Words, Sentences, Discourse. *Discourse Processes*, vol. (25) : 211-257.

Choi F. (2000). Advances in domain independent linear text segmentation. In *Proceedings of NAACL'00* : 26-33.

Choi F., Wiemer-Hastings P. et Moore J. (2001). Latent Semantic Analysis for Text Segmentation. In *Proceedings of NAACL'01* : 109-117.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol. (41) : 391-407.

Degand L. et Bestgen Y. (2003). Towards automatic retrieval of idioms in French Newspaper Corpora. *Literary and Linguistic Computing,* vol. (18) : 249-259.

Dumais S.T. (1995). Latent semantic indexing (LSI): TREC-3 report. In Harman D. (Ed.), *Proceedings of The 3rd Text Retrieval Conference (TREC-3)* : 219-230.

Ferret O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of COLING 2002* : 260-266.

Foltz P.W., Kintsch W. et Landauer T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, vol. (25) : 285-307.

Gernsbacher M.A. (1990). *Language comprehension as structure building*. LEA.

Kintsch W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, vol. (7) : 257-266.

Kintsch W. (2001). Predication. *Cognitive Science*, vol. (25) : 173-202.

Klebanov B. et Wiemer-Hastings P. (2002). Using LSA for Pronominal Anaphora Resolution. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing*.

Landauer T.K. et Dumais S.T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, vol. (104) : 211–240.

Landauer T.K., Foltz P.W. et Laham D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, vol. (25) : 259-284.

Lebart L. et Salem A. (1992) *Statistique textuelle.* Dunod.

Lebart L., Morineau A. et Piron M. (2000). *Statistique exploratoire multidimensionnelle* (3ième édition). Dunod.

Lemaire B., Bianco M., Sylvestre E. et Noveck I. (2001). Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente. In Paugam Moisy H., Nyckees V. et Caron-Pargue J. (Eds), *La cognition entre individu et société* (actes du colloque de l'ARCo). Hermès : 309-320.

Manning C.D. et Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Rajman M. et Besançon R. (1997). Text Mining : Natural Language Techniques and Text Mining Applications. In *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics*. Chapam & Hall.

Riloff E. (1995). Little words can make a big difference for dext classification. In *Proceedings of the 18th Annual International ACM SIGIR. Conference on Research and Development in Information Retrieval* : 130-136.

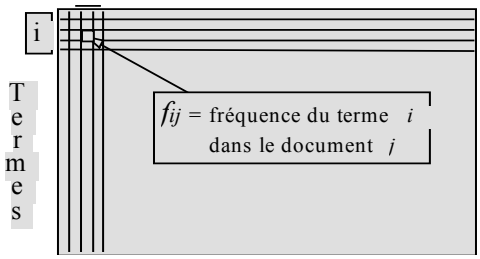
Schmidt H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Version électronique disponible sur [http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html].

Utiyama M. et Isahara H. (2001). A Statistical Model for Domain-Independent Text Segmentation. In *Proceedings of ACL'2001* : 491-498.

Wolf M.B.W., Schreiner M.E., Rehder B. et Laham D. (1998). Learning from text: Matching readers and texts by Latent Semantic Analusis. *Discourse Processes,* vol (25) : 309-336.

Annexe 1 : Les étapes d'une analyse sémantique latente

1) Obtention d'un tableau lexical
« termes * documents »
(nombre d'occurrences de chaque
terme dans chaque document)



2) Transformation des fréquences
afin de privilégier les termes
les plus informatifs

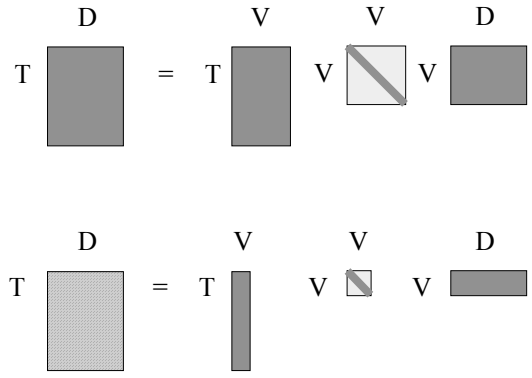
Transformation des fréquences
afin de privilégier les mots les plus informatifs

$$f_{ij}' = \frac{\log(f_{ij}+1)}{-\sum_j \frac{f_{ij}}{\sum_j f_{ij}} \log\left(\frac{f_{ij}}{\sum_j f_{ij}}\right)}$$

Pondération locale :
réduit l'impact des mots
très fréquents dans un document

Pondération globale :
réduit l'impact des mots « peu informatifs », c'est-à-dire qui
apparaissent avec une fréquence constante dans les documents
Exemples : mots peu informatifs : faire, pouvoir, deux, nouveau
mots très informatifs : neutrino, cancérogène, fiacre

3) Décomposition en valeurs singulières
- Compression de l'information par la sélection des k dimensions orthogonales les + importantes ($100 \leq k \leq 300$)
- Permet d'obtenir les vecteurs qui représentent les termes dans l'espace comprimé



4) Emploi :
- Calculer la proximité sémantique entre des mots ou des segments
- Le sens d'un mot est représenté par un vecteur
- La similarité entre deux mots est mesurée par le cosinus entre les vecteurs correspondants (idem pour les segments)