Limites de la lemmatisation pour l'extraction de significations

Benoît Lemaire

TIMC-IMAG (CNRS UMR 5525) - Université Grenoble 1
38 607 La Tronche Cedex - France

Abstract

Corpus lemmatization is a widely used procedure which is sometimes done for the sake of following a tradition. This paper highlights the limits of this process in the case of automatic extraction of semantic information, that is, when the context in which words occur is used. First, we uncovered significant differences between contexts of singular and plural forms of 58 nouns in a large French corpus. Systematically replacing plural forms by singular forms might therefore affect the performances of semantic extraction systems. Then, we relied on Latent Semantic Analysis to show in another way that the two contexts are different and that LSA performances on a vocabulary test decrease when the corpus is lemmatized. Lemmatizing corpora for such a usage might therefore work against the general intention.

Résumé

La lemmatisation des corpus est une procédure répandue que l'on effectue parfois par simple respect d'une tradition. Cet article met en évidence les limites de cette opération dans le cas de l'extraction automatique d'informations sémantiques, c'est-à-dire lorsque le contexte d'apparition des mots est utilisé. Nous montrons dans une première partie que les contextes des formes plurielles et singulières de 58 mots dans un vaste corpus diffèrent significativement, ce qui laisse penser que remplacer les uns par les autres peut affecter les performances des systèmes d'extraction de significations. Dans une seconde partie, nous recourons à l'analyse de la sémantique latente (LSA) pour montrer d'une autre manière que les contextes des deux formes ne sont pas les mêmes et que les performances du système sur un test de vocabulaire diminuent dès lors que le corpus est lemmatisé. Le lemmatisation des corpus pour un tel usage va donc peut-être à l'encontre du but recherché.

Mots-clés : lemmatisation, contexte, cotexte, analyse de la sémantique latente, LSA, corpus.

1. Introduction

La lemmatisation des corpus, qui consiste à remplacer chaque mot par sa forme canonique, est une opération courante dont les avantages et les inconvénients ont fait l'objet de nombreux articles (Brunet, 2002 ; Xu & Croft, 1998) et débats mémorables comme ceux opposant Muller et Tournier (Kastberg Sjöblom, 2004). Parmi les avantages, on peut citer la réduction du nombre de formes à considérer et l'augmentation des occurrences de chaque forme dans le corpus. Cette considération était cruciale il y a quelques années encore, lorsque les mémoires de nos ordinateurs ne pouvaient traiter de volumineux corpus ou de gigantesques matrices. Elle le reste cependant lorsque l'on ne dispose que de petits corpus spécialisés pour lesquels on veut augmenter les fréquences des formes. Les inconvénients de la lemmatisation proviennent de la perte d'informations résultant du remplacement d'un mot par son lemme. Notre hypothèse est que cette perte d'informations est préjudiciable aux algorithmes qui utilisent le contexte des mots (ou, linguistiquement parlant, le cotexte) puisque ce dernier n'est pas nécessairement le même selon la forme des mots. Par exemple, le contexte du mot *soleil* n'est pas le même que celui du mot *soleils*, notamment parce que les prédicats associés

à ces deux formes ne sont pas systématiquement les mêmes. Ainsi, le mot *brille* apparaît dans un de nos corpus généraux 39 fois avec la forme *soleil* et jamais avec la forme *soleils*. De même, le mot *rayon* apparaît 311 fois avec *soleil* et seulement 2 fois avec *soleils*. A l'inverse, le mot *étoiles* y apparaît 68 fois avec *soleils* et 10 fois avec *soleil*.

Ce sont donc les différences de contextes entre les diverses formes d'un lemme qui défavoriseraient les corpus lemmatisés, pour des algorithmes qui utilisent justement ce contexte. Pour étayer cette hypothèse, nous avons étudié les contextes de 58 noms, dans leur forme singulière ou plurielle, au sein d'un volumineux corpus. Nous avons ensuite eu recours à l'analyse de la sémantique latente, une méthode qui utilise largement le contexte dans lequel les mots apparaissent pour en extraire des informations sémantiques, afin d'étudier les différences de contextes entre ces différentes formes et les performances des différentes versions d'un corpus.

2. Comparaison des distributions des cooccurrents

Ce premier test vise à comparer les contextes des formes singulières et plurielles d'une liste de 58 noms. Plus précisément, il s'agit de comparer les distributions des cooccurrents de chaque forme, c'est-à-dire les mots qui apparaissent conjointement à cette forme dans le corpus. Pour cela, nous avons utilisé le corpus Corpatext 1.02 de la Wordthèque, disponible sur le site http://lexique.org. Ce corpus général est composé de 2 700 ouvrages en français totalisant 36 millions de mots. Nous avons sélectionné 58 noms français dont la forme singulière différait de la forme plurielle, ayant de hautes fréquences dans la langue (grâce aux données issues du site http://lexique.org) tout en éliminant quelques mots fortement polysémiques (comme *peine* ou *fond*). Pour chacun de ces noms, nous avons calculé l'ensemble des mots qui apparaissent conjointement avec eux, à une distance de six mots maximum à droite ou à gauche, l'unité étant le paragraphe. Dans la phrase précédente, les cooccurrents de *distance* sont donc [*apparaissent, conjointement, avec, eux, à, une, de, six, mots, maximum, à, droite*]. Pour chaque mot, nous obtenons donc une distribution des cooccurrents que nous comparons avec la distribution des cooccurrents de sa forme plurielle par une mesure de corrélation, en exprimant chaque distribution dans l'union des mots des deux distributions. Ces valeurs de corrélation sont indiquées dans la deuxième colonne du tableau 1. On peut noter que cette valeur de corrélation est très faible pour certains mots comme *soir* ou *visage*, ce qui laisse penser que les contextes de leurs formes singulières et plurielles sont très différents.

Afin de situer ces valeurs de corrélation par rapport à une norme, nous avons également calculé les corrélations entre les distributions des cooccurrents des formes singulières dans une moitié de corpus et dans l'autre moitié. Pour éviter un biais lié à la répartition des ouvrages dans le corpus, nous n'avons pas simplement coupé le corpus en deux, mais nous avons placé dans la première moitié les paragraphes de numéros impairs et dans la seconde moitié les paragraphes de numéros pairs. La troisième colonne du tableau 1 présente ces valeurs.

Tableau 1 : Valeurs des corrélations entre les distributions des cooccurrents des formes singulières et plurielles (corpus entier) ou des formes singulières sur deux moitiés de corpus pour 58 noms. Les corrélations supérieures à 0,8 sont sur un fond gris foncé et les corrélations entre 0,6 et 0,8 sont sur un fond gris clair.

MOT	Sing. / pluriel (corpus entier)	Sing. (moitié 1) / sing. (moitié 2)	MOT	Sing. / pluriel (corpus entier)	Sing. (moitié 1) / sing. (moitié 2)
air	.67	.98	matin	.14	.80
amour	.07	.92	mer	-.01	.99
an	.24	.57	monde	-.01	.99
bouche	-.03	.98	mort	.02	.98
bout	.00	.79	mot	.04	.93
bruit	.50	.93	nom	.06	.97
chambre	.04	.92	nuit	.46	.99
cheveu	.42	.99	oeil	.02	.83
chose	.30	.96	peur	.09	.72
ciel	.08	.91	pied	.50	.98
coeur	.69	.83	plaisir	.04	.88
coup	.37	.99	porte	.09	.99
dieu	.37	.98	question	-.02	.31
doute	.02	.89	regard	.02	.99
eau	.00	.24	route	.26	.98
enfant	.67	.98	rue	-.01	.97
famille	.00	.82	sang	-.01	.89
femme	.55	.98	silence	.03	.64
fille	.64	.81	soir	-.01	.97
fin	.00	.77	soleil	-.02	.97
force	.23	.85	table	.00	.85
guerre	.05	.97	tour	.00	.99
heure	.22	.90	train	-.10	.98
histoire	.28	.98	travail	.12	.93
homme	.48	.98	vent	-.01	.86
instant	.37	.97	vie	.17	.54
lit	.22	.62	ville	.01	.98
main	.71	.99	visage	-.01	.95
maison	.04	.16	voiture	.23	.95

Il apparaît clairement que les corrélations pour un même mot dans deux moitiés de corpus sont très élevées et toujours supérieures aux corrélations pour les deux formes d'un mot. Ce premier test nous conforte dans l'idée que les contextes des formes singulières et plurielles peuvent être relativement éloignés.

3. Voisinages sémantiques avec et sans lemmatisation

Une autre manière de comparer les contextes des formes singulières et plurielles consiste à recourir à une méthode induisant des similarités sémantiques à partir de ces contextes (et exclusivement à partir de ces contextes) et à comparer les similarités produites. Si les contextes étaient similaires, on devrait obtenir des mesures de similarités proches.

L'analyse de la sémantique latente (ou LSA, pour Latent Semantic Analysis) (Landauer, 2002) est une telle méthode. A partir d'une matrice M d'occurrences des mots dans les paragraphes d'un corpus brut, LSA y applique une décomposition en valeurs singulières : $M=U\Sigma V^T$, puis annule les plus petites valeurs singulières de la matrice diagonale Σ, de façon à supprimer le bruit inhérent au corpus et à ne conserver que les 300 dimensions les plus importantes. Cette valeur empirique correspond au maximum de performances sur lequel semblent s'accorder les résultats de la littérature (Landauer & Dumais, 1997). Chaque mot et chaque paragraphe sont ainsi représentés par un vecteur dans un espace à 300 dimensions. Il est alors possible de calculer une similarité sémantique entre mots ou entre paragraphes, en calculant le cosinus des angles de leurs vecteurs respectifs. D'autres mesures de similarités entre vecteurs sont possibles, mais c'est généralement le cosinus qui est utilisé.

Un vecteur peut être associé à tout nouveau document par simple somme des vecteurs de ses mots. De nombreux tests dans la littérature montrent que les similarités obtenues par LSA s'accordent relativement bien avec les jugements de similarités des humains, à la fois au niveau des mots isolés que des phrases ou des textes (Wolfe et al., 1998, Howard & Kahanna, 2007). La procédure mathématique qui sous-tend LSA ainsi que de nombreuses applications sont publiées dans un récent ouvrage édité par Landauer et al. (2007).

Notre objectif est de comparer les voisins sémantiques des formes plurielles et singulières de plusieurs mots afin de mettre au jour d'éventuelles différences qui ne pourraient être dues qu'aux différences de contexte entre les deux formes. Le principe est cependant quelque peu différent du test précédent, dans lequel nous comparions les contextes bruts. En effet, LSA ne se base pas uniquement sur les cooccurrences (le contexte brut) pour établir les similarités puisqu'il est capable de considérer deux mots similaires qui n'apparaissent jamais ensemble dans aucun paragraphe (Lemaire & Denhière, 2006). LSA implémente en fait l'idée que deux mots sont sémantiquement similaires s'ils apparaissent dans des paragraphes similaires (pas nécessairement identiques). Des paragraphes similaires contiennent des mots similaires (pas nécessairement identiques).

Pour tester notre hypothèse, nous avons donc eu recours à un corpus général de 13 millions de mots. Ce corpus contient 5 millions de mots issus du quotidien Le Monde, 5 millions de mots provenant de romans ou d'articles scientifiques et 3 millions de mots issus de textes pour enfants et adolescents. Ce corpus a été traité par LSA, qui a construit et réduit à 300 dimensions une matrice de 100.983 mots par 189.424 paragraphes. Nos travaux antérieurs (Denhière & Lemaire, 2004) ont montré que ce corpus était suffisant pour que des similarités sémantiques pertinentes puissent en être extraites. Notons que le contexte est ici le paragraphe et non une fenêtre glissante comme dans la section précédente.

Pour choisir une liste de mots pertinente, nous avons utilisé le poids que LSA attribue à chaque mot. Cette valeur dépend de sa fréquence dans les paragraphes et dans le corpus : les mots qui apparaissent fréquemment dans un grand nombre de paragraphes ont un poids faible et les mots qui apparaissent peu fréquemment et dans quelques paragraphes seulement ont un poids élevé. Nous avons utilisé cette valeur pour sélectionner 39 mots allant des fréquences élevées (poids de 0.3) aux fréquences plus rares (poids de 0.7). Au-delà de 0.7, l'expérience montre qu'on se trouve face à des mots moins fréquents pour lesquels les mesures de similarités de LSA sont moins fiables. Nous avons choisi, pour chaque centième de poids à partir de 0.3, le premier nom qui n'était pas une forme verbale, dont la forme plurielle était orthographiée différemment et de poids inférieur à 0.7. Ce dernier critère visait à éviter des mots dont la forme singulière aurait été fréquente, mais la forme plurielle rare. Nous disposons donc d'une liste de 39 mots couvrant une large gamme de fréquences.

Pour chacun de ces mots, nous avons recherché le rang de sa forme plurielle dans la liste de ses voisins sémantiques produits par LSA (en excluant de la liste les mots peu fréquents, dont le poids, là encore, était supérieur à 0,7). Chaque fois que cette valeur n'est pas 1, c'est-à-dire chaque fois que le voisin le plus proche d'un mot n'est pas sa forme plurielle, cela signifie donc qu'il existe des mots dont le contexte est statistiquement plus proche de la forme singulière que celui de la forme plurielle ne l'est de la forme singulière. Le tableau 2 présente les résultats. Par exemple, le premier voisin sémantique du mot *pied* est le mot *pieds*. Cela signifie qu'il n'y a pas de mots qui ont un contexte plus proche de *pied* que le mot *pieds*. En revanche, le mot *présidents* n'apparaît que comme le 22e voisin de *président*. Il y a donc 21 mots qui ont un contexte statistiquement plus proche de celui de *président*. Remplacer systématiquement *présidents* par *président* modifiera donc le contexte de ce mot. La seconde colonne indique le premier voisin, celui qui est, dans ce corpus, plus proche de la forme singulière que ne l'est la forme plurielle. Dans certains cas, la forme plurielle n'apparaît pas parmi les 500 premiers voisins sémantiques, ce qui est le signe d'une différence manifeste de contextes.

4. Performances avec et sans lemmatisation

Pour conclure cette étude, nous avons effectivement lemmatisé un corpus et comparé les performances des versions lemmatisée et non lemmatisée traitées par LSA sur un test de vocabulaire. Le corpus en question est composé des textes pour enfants du corpus précédent, pour un total de 3 millions de mots. Ce corpus est adapté pour notre test de vocabulaire, antérieurement conçu par Denhière et Legros (non publié) et destiné aux enfants du CE1 au cm2. Ce test contient 120 items, chacun composé d'un mot et de quatre définitions : la bonne définition du mot, une définition proche, une définition éloignée et une définition sans rapport. La tâche du participant est d'identifier la bonne définition parmi les quatre. Voici deux exemples d'items :

accompagner :

- ❒ passer devant
- ❒ regarder attentivement pendant longtemps en admirant
- ❒ aller avec quelqu'un quelque part
- ❒ guider

épaule :

- ❒ manteau sans manche qui couvre le corps et les bras
- ❒ partie du corps située en haut du dos
- ❒ petite bosse que les hommes ont à l'avant du cou
- ❒ endroit où le bras s'attache au corps

Tableau 2 : Premiers voisins sémantiques et rangs de la forme plurielle de 39 mots de fréquences décroissantes

MOT	premier voisin	rang de la forme plurielle	MOT	premier voisin	rang de la forme plurielle
président	république	22	fleur	fleurs	1
enfant	gâté	>500	drame	tragique	8
état	actuel	19	attentat	revendiqué	4
guerre	golfe	5	scène	metteur	12
mère	maternel	9	restaurant	frites	75
mot	étymologie	23	successeur	nomination	>500
ami	amitié	2	filet	plat	4
école	élèves	8	proportion	moyenne	21
développement	développer	187	écureuil	hérisson	12
pied	pieds	1	créateur	improvisation	126
bruit	bruits	1	satellite	câble	23
expression	exprimer	4	sociologue	sociologique	6
réponse	question	6	batterie	jazz	>500
mémoire	mémoires	1	entrepreneur	factures	66
élève	lycée	3	prairie	herbe	20
salarié	salariés	1	moulin	vent	495
zone	frontière	11	rédacteur	chef	>500
troupe	scène	403	millimètre	globules	24
lecteur	roman	8	poire	poires	1
tour	eiffel	7			

Pour chaque item, nous avons calculé le cosinus entre le vecteur correspondant au mot-cible et le vecteur correspondant à chacune des définitions (défini comme la somme des vecteurs des mots qui la composent), puis sélectionné la définition correspondant au cosinus maximum. La version non lemmatisée du corpus traitée par LSA obtient un score de 53% de réponses correctes, ce qui est bien au-dessus du hasard. Pour information, les enfants de CM2 ont un score moyen de 72% et les enfants de CE1 ont le même score de 53%. LSA appliqué à notre corpus a résolu correctement les deux exemples précédents (réponse 3 pour *accompagner* et 4 pour *épaule*).

La version lemmatisée du corpus a été réalisée avec Flemm (Namer, 2000) après étiquetage par le catégoriseur de Brill sur les fichiers français distribués par l'ATILF. Le test a évidemment été lemmatisé lui aussi et contrôlé manuellement. Après traitement par LSA, nous avons obtenu un score de 44%, ce qui est moins bon que la version non lemmatisée. Par exemple, cette version échoue à identifier la définition correcte de *épaule* et lui préfère la définition éloignée. La lemmatisation totale de ce corpus diminue donc les performances de LSA sur notre corpus.

Il est possible que la lemmatisation des verbes soit moins préjudiciable à de telles méthodes basées sur le contexte puisqu'on peut faire l'hypothèse qu'il y a moins de variations dans le contexte des différentes formes fléchies des verbes. Les formes *coupa*, *couperait, couperons, couperiez,* etc. apparaissent probablement dans des contextes assez proches, bien qu'il soit possible qu'une analyse plus fine mette au jour des différences de contexte. Pour tester cette hypothèse, nous avons effectué une nouvelle version du corpus dans lequel seuls les verbes étaient lemmatisés. Le nombre de réponses correctes est alors remonté à 49%, ce qui est meilleur qu'avec la version totalement lemmatisée.

Ces derniers tests mériteraient bien sûr d'être répliqués sur d'autres corpus, mais ils laissent penser qu'il est peut-être opportun de ne lemmatiser que partiellement les corpus.

5. Conclusion

Ce travail ne s'intéresse qu'à la lemmatisation dans le cadre de l'extraction des significations à partir du contexte. Il n'est donc pas sensible à l'argument selon lequel ne pas lemmatiser revient à compter ensemble torchons et serviettes (Mayaffre, 2005), c'est-à-dire à considérer de la même manière les homographes. En effet, la prise en compte du contexte par des algorithmes d'extraction des significations comme LSA leur permet de faire cette désambiguïsation. Inutile donc de distinguer préalablement *partis* (nom pluriel) et *partis* (participe passé) puisqu'ils n'apparaissent pas dans les mêmes contextes. La méthode leur assignera des voisins sémantiques dans des parties de l'espace bien distinctes.

Plusieurs travaux s'interrogent sur les limites de la lemmatisation. Il ne semble pas cependant qu'un consensus se dégage clairement. Beaucoup de travaux sont réalisés sur la langue anglaise, mais il est clair que les résultats dépendent des caractéristiques de la langue étudiée. Dans le domaine de la classification des textes, on peut noter les travaux de Riloff (1995) qui montrent clairement que la lemmatisation modifie les résultats. Dans le domaine de la recherche d'informations, et plus spécifiquement de la réponse à des questions, Billoti et al. (2004) aboutissent aussi à des résultats moins bons dès lors que les corpus sont lemmatisés. Ceci contredit cependant les travaux de Monz (2003) qui montrent une amélioration systématique de la lemmatisation des corpus en termes de rappel et de précision. LSA a également été utilisée pour corriger des mots de manière contextuelle par Jones et Martin (1997) qui affirment avoir obtenu de meilleurs résultats avec lemmatisation préalable, mais sans communiquer de données précises sur cet effet. Enfin, Zipitria et al. (2006) ne remarquent aucune différence entre un corpus basque non lemmatisé et sa version lemmatisée, tous deux traités par LSA, pour mesurer la cohérence et la compréhension des textes chez des apprenants.

Nous avons ici tenté de nous situer quelque peu en amont de ces travaux, en n'étudiant pas uniquement les performances des systèmes, mais d'abord les contextes dans lesquelles les différentes formes des mots apparaissent. Il resterait à répliquer comme d'habitude ces résultats et à réaliser d'autres études, notamment au niveau des verbes. Quoi qu'il en soit, nous pensons que la lemmatisation n'est probablement pas une opération anodine et qu'il convient d'être prudent dès lors que les mots des corpus sont étudiés en fonction de leur contexte.

Remerciements

Nous tenons à remercier Guy Denhière qui a largement contribué à la constitution de deux des corpus présentés dans cet article.

Références

Brunet E. (2002). Le lemme comme on l'aime. In *Actes des 6es Journées Internationales d'Analyse Statistique des Données Textuelles*, 221-232.

Denhière G., Lemaire B. (2004). A Computational Model of a Child Semantic Memory. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004)*, 297-302.

Howard M. W. & Kahanna M. J. (2007). Semantic Structure and Episodic Memory. In T.K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (Eds), *Handbook of Latent Semantic Analysis.* Mahwah, Lawrence Erlbaum Associates.

Jones M. J. & Martin J. H. (1997). Contextual spelling correction using Latent Semantic Analysis. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*.

Kastberg Sjöblom M. (2004). L'écriture de J.M.G. Le Clézio, une approche lexicométrique. *Texto !.*

Landauer T. K., McNamara D. S., Dennis S., Kintsch W. (Eds) (2007). *Handbook of Latent Semantic Analysis.* Mahwah, Lawrence Erlbaum Associates.

Lemaire B., Denhière G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current Psychology Letters* 18(1).

Mayaffre D. (2005). De la lexicométrie à la logométrie. *L'Astrolabe. Recherche littéraire et Informatique* (revue électronique).

Namer F. (2000). Flemm: Un analyseur Flexionnel de Français à base de règles. In C. Jacquemin (Ed), *Traitement automatique des Langues pour la recherche d'information*. Paris, Hermes, 523-47.

Riloff E. (1995). Little Words can Make a Big Difference for Text Classification. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 130-136.

Xu J. & Croft B. (1998). Corpus-Based Stemming using Cooccurrence of Word Variants. *ACM Transactions on Information Systems* 16(1), 61-81.

Wolfe M. B. W., Schreiner M. E., Rehder B., Laham D., Foltz P. W., Kintsch W., Landauer T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, 25, 309-336.

Zipitria I., Arruarte A. & Elorriaga J. A. (2006). Observing lemmatization effect in LSA coherence and comprehension grading of learner summaries. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, Lecture Notes in Computer Science 4053, 595-603, Berlin, Springer.