Amélioration de la segmentation automatique des textes grâce aux connaissances acquises par l'analyse sémantique latente.

Yves Bestgen

Centre pour l'étude du texte et du discours – PSOR - Université catholique de Louvain

Place du Cardinal Mercier, 10 - B1348 Louvain-la-Neuve Belgique

yves.bestgen@psp.ucl.ac.be

Mots-clés : Segmentation automatique de textes, Analyse sémantique latente (ASL)

Keywords: Automatic text segmentation, Latent semantic analysis (LSA)

Résumé Choi, Wiemer-Hastings et Moore (2001) ont proposé d'employer l'analyse sémantique latente (ASL) pour extraire des connaissances sémantiques à partir de corpus afin d'améliorer l'efficacité d'un algorithme de segmentation des textes. En comparant l'efficacité du même algorithme selon qu'il prend en compte des connaissances sémantiques complémentaires ou non, ils ont pu montrer les bénéfices apportés par ces connaissances. Dans leurs expériences cependant, les connaissances sémantiques avaient été extraites d'un corpus qui contenait les textes à segmenter dans la phase de test. Si cette hyperspécificité du corpus d'apprentissage explique la plus grande partie de l'avantage observé, on peut se demander s'il est possible d'employer l'ASL pour extraire des connaissances sémantiques génériques pouvant être employées pour segmenter de nouveaux textes. Les deux expériences présentées ici montrent que la présence dans le corpus d'apprentissage du matériel de test a un effet important, mais également que les connaissances sémantiques génériques dérivées de grands corpus améliorent l'efficacité de la segmentation.

Abstract Choi, Wiemer-Hastings and Moore (2001) proposed to use latent Semantic Analysis to extract semantic knowledge from corpora in order to improve the accuracy of a text segmentation algorithm. By comparing the accuracy of the very same algorithm depending on whether or not it takes into account complementary semantic knowledge, they were able to show the benefit derived from such knowledge. In their experiments, semantic knowledge was, however, acquired from a corpus containing the texts to be segmented in the test phase. If this hyper-specificity of the training corpus explains the largest part of the benefit, one may wonder if it is possible to use LSA to acquire generic semantic knowledge that can be used to segment new texts. The two experiments reported here show that the presence of the test materials in the training corpus has an important effect, but also that the generic semantic knowledge derived from large corpora clearly improves the segmentation accuracy.

1 Améliorer la segmentation des textes par l'adjonction de connaissances sémantiques complémentaires?

Pendant les dix dernières années, de nombreuses méthodes ont été proposées pour segmenter automatiquement des textes en fonction des thèmes qui les composent sur la base de la cohésion lexicale. La distinction principale entre ces méthodes réside dans le contraste entre les approches basées exclusivement sur l'information contenue dans le texte à segmenter comme la cohésion lexicale (par exemple, Choi, 2000 ; Hearst, 1997 ; Heinonen, 1998 ; Kehagias, Pavlina, Petridis, 2003 ; Utiyama, Isahara, 2001), et celles qui reposent sur des connaissances sémantiques complémentaires extraites de dictionnaires et de thesaurus (par exemple, Kozima 1993 ; Lin, Nunamaker, Chau, Chen, 2004 ; Morris, Hirst, 1991), ou des collocations observées dans de grands corpus (Bolshakov, Gelbukh 2001 ; Choi *et al.*, 2001 ; Ferret, 2002 ; Kaufmann, 1999 ; Ponte, Croft, 1997). Selon leurs auteurs, les méthodes qui utilisent des connaissances supplémentaires apportent une réponse aux problèmes posés par les phrases qui relèvent du même thème tout en ne partageant aucun mot commun ou par la présence de synonymes et d'hyperonymes. Des arguments empiriques en faveur de ces méthodes ont été récemment présentés par Choi *et al.* (2001) dans une étude basée sur l'analyse sémantique latente (ASL : Latent semantic analysis, Latent semantic indexing, Deerwester *et al.,* 1990), une technique statistique d'extraction d'espaces sémantiques à partir de corpus qui permet l'estimation de la similarité sémantique entre des mots, des phrases ou des paragraphes. En comparant l'efficacité du même algorithme selon qu'il prend en compte ou non ces connaissances sémantiques complémentaires, Choi *et al.* (2001) ont mis en évidence l'avantage dérivé de telles connaissances.

Toutefois, les implications de l'étude de Choi *et al.* pour la segmentation des textes et, plus généralement, pour l'utilisation de l'ASL dans le traitement automatique du langage sont rendues peu claires en raison de la méthodologie qu'ils ont employée. Dans leurs expériences, les connaissances sémantiques ont été extraites d'un corpus qui contenait les textes qui ont été segmentés dans la phase de test. On peut donc se demander si la plus grande partie des bénéfices obtenus par l'ajout de connaissances sémantiques n'est pas due à cette hyperspécificité du corpus d'apprentissage (c.-à-d. inclure le matériel de test). Si c'est le cas, cela met en question la possibilité d'employer l'ASL pour extraire des connaissances sémantiques génériques pouvant être utilisées pour segmenter de nouveaux textes. A priori, le problème ne semble pas très important, parce que Choi *et al.* ont utilisé un grand nombre de petits échantillons de test pour évaluer leur algorithme, chaque échantillon ne représentant en moyenne que 0.15% du corpus d'apprentissage. La présente étude montre, toutefois, que la présence du matériel de test dans le corpus d'apprentissage a un effet important, mais également que les connaissances sémantiques génériques dérivées de grands corpus améliorent nettement l'efficacité de l'algorithme de segmentation. Cette conclusion est issue de deux expériences dans lesquelles la présence ou l'absence du matériel de test dans le corpus d'apprentissage pour l'ASL est manipulée. La première expérience est basée sur le matériel employé par Choi *et al.*, un petit corpus de 1.000.000 de mots. La deuxième expérience est basée sur un corpus beaucoup plus grand (25.000.000 mots). Avant de présenter ces expériences, l'algorithme et l'utilisation par Choi *et al.* de l'ASL dans ce cadre sont décrits.

2 Les deux versions de l'algorithme de Choi

L'algorithme de segmentation proposé par Choi (2000) se compose des trois étapes habituellement présentes dans les procédures de segmentation basées sur la cohésion lexicale. Premièrement, le document à segmenter est divisé en unités textuelles minimales, habituellement les phrases. Ensuite, un indice de similarité entre chaque paire d'unités prises deux par deux est calculé. Chaque valeur brute de similarité est réexprimée sous une forme ordinale en prenant la proportion de valeurs voisines qui sont plus petites qu'elle. Pour finir, le document est segmenté répétitivement selon les frontières entre les unités qui maximisent la somme des similarités moyennes à l'intérieur des segments ainsi constitués.

L'étape la plus intéressante pour la présente étude est celle qui calcule les similarités interphrases. La procédure initialement proposée par Choi (2000), C99, reposait exclusivement sur l'information contenue dans le texte à segmenter. Chaque phrase est représentée par un vecteur construit selon le modèle vectoriel classique (Manning, Schütze, 1999, pp. 539ff) et la similarité entre deux phrases est calculée au moyen de la mesure de cosinus entre les vecteurs correspondants. Dans une première évaluation basée sur la procédure décrite ci-dessous, Choi a montré que son algorithme était plus efficace que plusieurs autres approches telles que *TextTilling* (Hearst, 1994), *Segmenter* (Kan, Klavans, McKeown, 1998) et le *Maximum-probability segmentation algorithm* de Utiyama et Isahara (2001).

Choi *et al.* (2001) ont proposé d'améliorer la mesure de similarité inter-phrases en prenant en compte les proximités sémantiques entre les mots estimées sur la base de l'analyse sémantique latente (ASL). Brièvement, l'ASL s'appuie sur la thèse qu'il est possible d'estimer la similarité sémantique entre des mots en analysant les contextes dans lesquels ces mots apparaissent (Deerwester *et al.* 1990 ; Degand, Spooren, Bestgen, 2004 ; Landauer, Dumais 1997). La première étape d'une analyse sémantique latente consiste en la construction d'un tableau lexical contenant les fréquences d'occurrence de chaque mot dans chacun des documents, un document pouvant être une phrase, un paragraphe, un texte, ... Pour extraire les dimensions sémantiques, ce tableau lexical subit une décomposition en valeurs singulières, une sorte d'analyse factorielle qui extrait les dimensions orthogonales les plus importantes. Après cette étape, chaque mot est représenté par un vecteur de poids indiquant sa force d'association avec chacune des dimensions. Ceci permet de mesurer la proximité sémantique entre deux mots quelconques en utilisant, par exemple, la mesure de cosinus entre les vecteurs correspondants. La proximité entre deux phrases (ou toutes autres unités textuelles), même lorsque ces phrases ne font pas partir du corpus initial, peut être estimée en calculant un vecteur pour chacune de ces phrases -- correspondant à la somme pondérée des vecteurs des mots qui les composent -- et puis en calculant le cosinus entre ces vecteurs (Deerwester *et al.* 1990). Choi *et al.* (2001) ont montré que l'utilisation de cette procédure pour calculer les similarités inter-phrases produit des performances supérieures à celles enregistrées au moyen de la version précédente de l'algorithme (basé seulement sur la répétition de mots).

3 Expérience 1

Le but de cette expérience est de déterminer l'impact de la présence du matériel de test dans le corpus d'apprentissage de l'ASL sur les résultats obtenus par Choi *et al.* (2001). Est-ce que les connaissances sémantiques extraites d'un corpus qui n'inclut pas le matériel de test améliorent également l'efficacité de la segmentation ?

3.1 Méthode

Cette expérience est basée sur la méthodologie développée par Choi (2000). Cette méthodologie a été également employée par plusieurs auteurs pour évaluer l'efficacité de leur système de segmentation (Brants, Chen, Tsochantaridis, 2002 ; Ferret, 2002 ; Kehagias *et al.*, 2003 ; Utiyama, Isahara, 2001). La tâche consiste à retrouver les frontières entre des textes concaténés. Chaque échantillon de test est une concaténation de dix segments de textes. Chaque segment est composé des *n* premières phrases d'un texte aléatoirement choisi dans deux sous-sections du corpus de Brown. Pour l'expérience, j'ai utilisé le matériel de test le plus général proposé par Choi (2000) dans lequel la taille des segments dans chaque échantillon varie aléatoirement de 3 à 11 phrases. Il est composé de 400 échantillons.

L'expérience vise à comparer l'efficacité de l'algorithme selon que le matériel de test est inclus dans le corpus d'apprentissage de l'ASL (condition *non autonome*) ou qu'il ne l'est pas (condition *autonome*). Un espace sémantique *non autonome*, correspondant à celui utilisé par Choi *et al.*, a été construit en utilisant l'entièreté du corpus de Brown comme corpus d'apprentissage. Quatre cents espaces *autonomes* différents ont été construits, un pour chaque échantillon de test, en retirant à chaque fois du corpus de Brown uniquement les phrases qui composent cet échantillon.

Pour extraire l'espace sémantique par l'ASL et pour appliquer l'algorithme de segmentation, une série de paramètres ont dû être fixés. Tout d'abord, les paragraphes ont été utilisés comme documents pour construire le tableau lexical parce que Choi *et al.* ont observé que de telles unités de taille moyenne étaient plus efficaces que des unités plus courtes comme les phrases. Les mots repris dans la liste de mots-outils (*stoplist*) de Choi ont été supprimés, ainsi que ceux qui n'apparaissaient qu'une seule fois dans l'ensemble du corpus. Les mots n'ont pas été tronqués en fonction de leur racine (*stemming*), suivant en cela la procédure de Choi *et al.* (2001). Pour établir l'espace sémantique, la décomposition en valeurs singulières a été réalisée par le programme SVDPACKC (Berry, 1992 ; Berry *et al.*, 1993), et les 300 premiers vecteurs singuliers ont été conservés. En ce qui concerne l'algorithme de segmentation, j'ai utilisé la version dans laquelle le nombre de frontières à trouver est imposé et fixé ici à neuf. Un masque de 11 x 11 a été employé pour la transformation ordinale, comme recommandé par Choi (2000).

3.2 Résultats

L'efficacité de la segmentation a été évaluée au moyen de l'indice utilisé par Choi *et al.* (2001) : le taux Pk d'erreur de segmentation (Beeferman, Berger, Lafferty, 1999) qui indique la proportion de phrases qui sont incorrectement classées comme appartenant au même segment ou incorrectement classées comme appartenant à des segments différents.

Les résultats[1] sont présentés dans la Figure 1. Les espaces autonomes donnent lieu à des performances plus faibles que l'espace non autonome, comme le confirme le test *t* pour échantillon appareillé (chaque échantillon de test étant utilisé comme une observation) qui est

[1] Ce taux d'erreur est en fait légèrement meilleur que celui obtenu par Choi et al. (2001), la différence pouvant être due à plusieurs facteurs tels que le prétraitement du corpus de Brown (identification des mots et des paragraphes) ou la fonction de pondération appliquée aux fréquences brutes, qui était ici la formule de pondération décrite dans Landauer, Foltz, et Laham (1998).

significatif pour un alpha plus petit que 0.0001. L'algorithme C99, qui n'utilise pas l'ASL pour estimer les similarités entre les phrases, produit un Pk de 0.13 (Choi *et al.*, 2001, tableau 3, ligne 3 : *no stemming*). Il s'avère donc que, si la condition *autonome* est meilleure que C99, l'avantage est très faible.

	Pk
Non autonome	0.084 (0.005)
Autonome	0.120 (0.006)

Figure 1 : Taux d'erreur et variance (entre parenthèses) pour les conditions non autonome et autonome.

Avant de conclure que la présence du matériel de test dans le corpus d'apprentissage de l'ASL a fortement modifié l'espace sémantique, une explication alternative doit être considérée. La perte d'efficacité en condition *autonome* pourrait être due au fait qu'il y a systématiquement légèrement moins de mots indexés dans les espaces sémantiques autonomes que dans l'espace *non autonome*. La suppression de chaque échantillon de test a entraîné la perte en moyenne de 23 mots différents sur un total de 25.847 mots qui sont indexés dans l'espace *non autonome*. Dans les espaces *autonomes*, ces mots ne sont pas disponibles pour estimer la similarité entre les phrases, tandis qu'ils sont utilisés dans l'espace *non autonome*. Afin de déterminer si ce facteur peut expliquer la différence d'efficacité, une analyse complémentaire a été effectuée sur l'espace *non autonome* dans laquelle, pour chaque échantillon de test, uniquement les mots présents dans l'espace *autonome* correspondant ont été pris en compte. De cette manière, seules les relations sémantiques peuvent jouer. Comparé à l'espace *non autonome* complet, je n'ai observé presque aucune diminution d'efficacité, le taux d'erreur Pk passant de 0.084 à 0.085 dans la nouvelle analyse. Ce résultat indique que ce ne sont pas les mots choisis pour le calcul des proximités qui importent, mais les relations sémantiques dans les espaces.

4 Expérience 2

L'expérience 1 a été menée sur le corpus d'apprentissage de Choi *et al.* (2001), un corpus de 1.000.000 de mots issus de textes de genres et de thèmes très différents. La petite taille du corpus et la diversité des textes pourraient avoir affecté les résultats de deux manières. D'abord, l'impact de la présence du matériel de test dans le corpus dépend probablement de ces caractéristiques du corpus. Retirer les premières phrases d'un texte devrait avoir moins d'effet si le corpus contient beaucoup de textes sur des thèmes similaires. En second lieu, un corpus plus volumineux permettrait probablement l'extraction d'un espace sémantique plus stable et plus efficace. Ceci pourrait produire une plus grande différence entre la version "ASL" de l'algorithme et celle qui n'utilise pas de connaissances sémantiques supplémentaires (C99). Pour ces raisons, une deuxième expérience a été menée sur la base d'un corpus beaucoup plus volumineux, comprenant les articles parus durant les années 1997 et 1998 dans le journal de langue française belge *Le Soir* (approximativement 52.000 articles et 26.000.000 mots). Dans ce corpus, chaque échantillon du matériel de test correspond en moyenne à 0.0066% du corpus complet. Cette deuxième expérience a également permis de comparer les conditions *non autonome* et *autonome* à une condition *antérieure* basée sur les articles parus dans le même journal, mais pendant les années 1995 et 1996 (approximativement 50.000 articles et plus de 22.000.000 mots). Cette condition nous informera à propos de la possibilité

d'employer l'ASL pour extraire des connaissances sémantiques plus génériques, puisque le corpus d'apprentissage de l'ASL est antérieur aux textes à segmenter. Il faut toutefois noter que ces connaissances étant extraites de la même source journalistique, les qualifier d'indépendantes seraient nettement excessifs.

4.1 Méthode

Le matériel de test a été extrait du corpus 1997-1998 suivant les directives données dans Choi (2000). Il se compose de 100 échantillons de dix segments, dont la longueur varie aléatoirement de 3 à 11 phrases. Trois types d'espace d'apprentissage pour l'ASL ont été construits. L'espace *non autonome* est basé sur l'entièreté du corpus 1997-1998. Cent espaces *autonomes* différents ont été construits comme décrit dans l'expérience 1. Enfin, un espace *antérieur* a été établi à partir du corpus 1995-1996. Les paramètres utilisés pour extraire les espaces sémantiques sont identiques à ceux employés dans l'expérience 1 sauf que, pour réduire la taille des tableaux lexicaux. les articles, et non les paragraphes, ont été utilisés comme documents et les mots ont été lemmatisés au moyen de TreeTagger (Schmid 1994).

4.2 Résultats

Globalement, les résultats sont similaires à ceux obtenus lors de la première expérience. Comme le montre la Figure 2, les espaces autonomes donnent lieu à des performances plus faibles que l'espace non autonome et, comme attendu, l'espace antérieur donne lieu à des performances encore plus faibles.

	Pk
Non autonome	0.074 (0.004)
Autonome	0.084 (0.005)
Antérieure	0.098 (0.005)

Figure 2 : Taux d'erreur et variance (entre parenthèses) pour les conditions non autonome, autonome et antérieure.

Toutefois, il est important de noter que l'algorithme C99, qui n'est pas basé sur l'analyse sémantique latente, produit un taux d'erreur Pk de 0.155, soit une valeur nettement plus mauvaise que celles obtenues avec les espaces *autonomes* (0.084) et avec l'espace *antérieur* (0.098). Ceci confirme l'utilité des connaissances sémantiques extraites de grands corpus pour estimer les similarités interphrases.

Par rapport à la première expérience, l'écart entre les conditions *non autonome* et *autonome* est beaucoup plus faible, passant de 0.036 pour l'expérience 1 à 0.01 pour l'expérience 2. Cet écart demeure néanmoins statistiquement significatif ($t(99) = 3.17$; $p = 0.002$). Bien que plus grand, l'écart entre les conditions *autonome* et *antérieure* (0.014), est statistiquement juste significatif ($t(99) = 2.04$; $p = 0.045$). La Figure 3 montre que la condition *autonome* surpasse la condition *antérieure* dans 46 échantillons, alors que l'inverse se produit dans 35 échantillons, les 19 échantillons restants ne montrant aucune différence entre ces deux conditions.

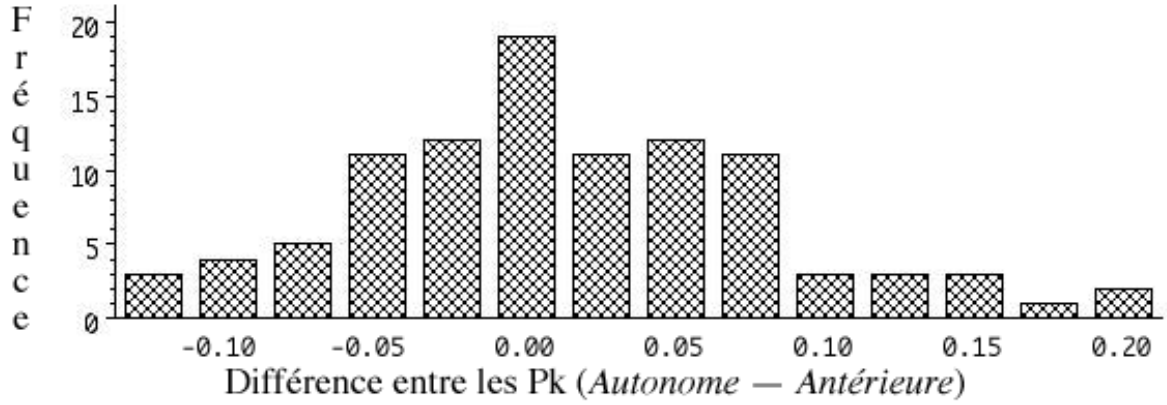

Figure 3 : Distribution des différences en Pk entre les conditions *Autonome* et *Antérieure*

On voit donc que l'avantage de la condition *autonome* sur la condition *antérieure* est principalement dû à quelques échantillons de test pour lesquels la condition *autonome* est nettement plus efficace. Rappelons également que la condition *antérieure* donne lieu à des résultats nettement meilleurs que ceux obtenus lorsque la segmentation s'effectue sans le recours à des connaissances sémantiques complémentaires.

5 Conclusion

Les deux expériences rapportées ici montrent que la présence du matériel de test dans le corpus d'apprentissage de l'ASL augmente l'efficacité de l'algorithme de segmentation même lorsqu'un corpus de plus de 25.000.000 mots est utilisé. Elles montrent également que l'utilisation de connaissances sémantiques indépendantes améliore l'efficacité de la segmentation et que ceci s'observe même lorsque ces connaissances sont extraites d'années antérieures de la même source. Cette observation souligne la possibilité de constituer par analyse sémantique latente des connaissances sémantiques plus ou moins génériques, c'est-à-dire, des connaissances qui peuvent être utilisées pour traiter de nouvelles données, comme cela a été récemment proposé dans la recherche de l'antécédent d'une anaphore, dans un système de reconnaissance de la parole ou en traduction automatique (Bellegarda, 2000 ; Klebanov, Wiemer-Hastings, 2002 ; Kim, Chang, Zhang, 2003). Une question à laquelle la présente étude ne répond pas concerne la possibilité d'utiliser un corpus tiré d'une autre source, telle qu'un autre journal. Bellegarda (2000) a observé, en reconnaissance automatique de la parole, qu'un tel espace sémantique est moins efficace. Cependant, évaluer la proximité sémantique entre deux phrases est probablement moins affecté par la source du corpus que de prédire le prochain mot d'un énoncé.

Récemment, plusieurs auteurs ont proposé des algorithmes de segmentation, basés principalement sur la programmation dynamique, qui égalent ou même surpassent les résultats de Choi (Ji, Zha, 2003, Kehagias *et al.*, 2003 ; Utiyama, Isahara, 2001). Ces algorithmes ne s'appuient pas sur des connaissances sémantiques supplémentaires. Les résultats de la présente étude suggèrent que leur efficacité pourrait encore être améliorée en prenant en compte de telles connaissances. Enfin, d'autres techniques que l'ASL ont été proposées pour extraire des connaissances sémantiques à partir de grands corpus tel pASL (Brants *et al.*, 2002). L'analyse sémantique latente étant relativement simple à mettre en pratique grâce à la disponibilité de

programmes très puissants tel que SVDPACKC (Berry *et al.*, 1993), son avantage principal est qu'elle est employée par une communauté de plus en plus large de chercheurs.

Une limitation importante de ce travail réside dans la tâche employée pour évaluer l'efficacité de l'algorithme de segmentation. Identifier les frontières entre des textes concaténés est une tâche artificielle et certainement moins complexe que de localiser les changements de thèmes à l'intérieur de textes. Le fait que la procédure et le matériel conçu par Choi soient devenus une sorte de "standard" employé par une série de chercheurs pour évaluer leur algorithme ne suffit pas à la légitimer. Il serait donc utile de confirmer les conclusions de la présente étude dans une tâche de segmentation intratexte.

Remerciements

Y. Bestgen est chercheur qualifié du Fonds National de la Recherche (FNRS). Cette recherche est financée par le projet FRFC n° 2.4535.02 et par une “Action de Recherche concertée” du Gouvernement de la Communauté française de Belgique.

Références

BEEFERMAN D., BERGER A., LAFFERTY J. (1999). Statistical models for text segmentation, *Machine Learning, Vol.* 34, pp. 177–210.

BELLEGARDA J. (2000). Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, pp. 78-84.

BERRY M. (1992). Large scale singular value computation. *International jornal of Supercomputer Application*, Vol. 6, 13-49.

BERRY M., DO T., O'BRIEN G., KRISHNA V., VARADHAN S. (1993). SVDPACKC: Version 1.0 user's guide, Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

BOLSHAKOV I., GELBUKH A. (2001). Text segmentation into paragraphs based on local text cohesion. In *Proceedings of Text, Speech and Dialogue (TSD-2001)*, 158–166.

BRANTS T., CHEN, F., TSOCHANTARIDIS I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of CIKM’02*, 211-218

CHOI F. (2000). Advances in domain independent linear text segmentation, In *Proceedings of NAACL-00*, 26–33.

CHOI F., WIEMER-HASTINGS P., MOORE J. (2001) Latent semantic analysis for text segmentation, In *Proceedings of NAACL’01*, 109–117.

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T., HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, pp. 391-407.

DEGAND L., SPOOREN W., BESTGEN Y. (2004). On the use of automatic tools for large scale semantic analyses of causal connectives. In *Proceedings of ACL 2004 Workshop on Discourse Annotation*, 25-32.

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

HEARST M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, Vol. 23, pp. 33–64.

HEINONEN O. (1998). Optimal multi-paragraph text segmentation by dynamic programming. In *Proceedings of 17th International Conference on Computational Linguistics (COLING-ACL'98)*, 1484-1486.

JI X., ZHA H. (2003). Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *Proceedings of SIGIR 2003*, 322-329.

KAN M., KLAVANS J., MCKEOWN K. (1998). Linear segmentation and segment significance. In *Proceedings of the 6th International Workshop of Very Large Corpora*, 197-205.

KAUFMANN, S. (1999). Cohesion and collocation: using context vectors in text segmentation. In *Proceedings of ACL'99*, 591–595.

KEHAGIAS A., PAVLINA F., PETRIDIS V. (2003). Linear text segmentation using a dynamic programming algorithm. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 171-178

KIM Y., CHANG J., ZHANG B. (2003). An empirical study on dimensionality optimization in text mining for linguistic. Knowledge Acquisition. In *Proceedings of PAKDD 2003*, 111–116.

KLEBANOV B., WIEMER-HASTINGS P. (2002). Using ASL for pronominal anaphora resolution. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing***,** 197-199.

KOZIMA H. (1993). Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 286-288.

LANDAUER T., DUMAIS S. (1997). A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review*, Vol. 104, pp. 211–240.

LANDAUER T., FOLTZ P., LAHAM D. (1998). An Introduction to latent semantic analysis. *Discourse Processes*, Vol. 25, pp. 259-284.

LIN M., NUNAMAKER J., CHAU, M., CHEN H. (2004). Segmentation of lecture videos based on text: A method combining multiple linguistic features. In *Proceedings of the 37th Hawaii International Conference on System Sciences*.

MANNING C., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

MORRIS J., HIRST G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* 17: 21-42.

PEVZNER L., HEARST M. (2002). A Critique and improvement of an evaluation metric for text segmentation, *Computational Linguistics,* Vol 28, pp. 19-36

PONTE J., CROFT W. (1997). Text segmentation by topic. *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries,* 120-129.

SCHMID H. (1994). Probabilistic Part-of-speech tagging using decision trees. P*roceedings of International Conference on New Methods in Language Processing* .

UTIYAMA M., ISAHARA H. (2001). A Statistical model for domain-independent text segmentation. *Proceedings of ACL'2001,* 491–498.