Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001)

Yves Bestgen*
FNRS - Université Catholique
de Louvain

Choi, Wiemer-Hastings, and Moore (2001) proposed to use Latent Semantic Analysis (LSA) 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 LSA 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 LSA corpus has an important effect, but also that the generic semantic knowledge derived from large corpora clearly improves the segmentation accuracy.

1. Improving Text Segmentation by Using Complementary Semantic Knowledge

For the last ten years, many methods have been proposed for the segmentation of texts in topically related units on the basis of lexical cohesion. The major distinction between these methods is in the contrast between the approaches based exclusively on the information contained in the text to be segmented, such as lexical repetition (e.g., Choi 2000; Hearst 1997; Heinonen 1998; Kehagias, Pavlina, and Petridis 2003; Utiyama and Isahara 2001), and those approaches that rest on complementary semantic knowledge extracted from dictionaries and thesauruses (e.g., Kozima 1993; Lin et al. 2004; Morris and Hirst 1991), or from collocations collected in large corpora (Bolshakov and Gelbukh 2001; Brants, Chen, and Tsochantaridis 2002; Choi et al. 2001; Ferret 2002; Kaufmann 1999; Ponte and Croft 1997). According to their authors, methods that use additional knowledge allow for a solution to problems encountered when sentences belonging to a unique topic do not share common words due to the use of hyperonyms or synonyms and allow words that are semantically related to be taken as positive evidence for topic continuity. Empirical arguments in favor of these methods have been provided recently by Choi et al. (2001) in a study using Latent Semantic Analysis (Latent Semantic Indexing, Deerwester et al. 1990) to extract a semantic space from a corpus allowing determination of the similarity of meanings of words, sentences, or paragraphs. By

* Center for Text and Discourse Studies, PSOR, Place du Cardinal Mercier 10, B-1348 Louvain-la-Neuve Belgium

comparing the accuracy of the very same algorithm according to whether or not it takes into account complementary semantic knowledge, they were able to show the benefit derived from such knowledge.

However, implications of Choi et al.'s study for text segmentation and for the use of LSA in natural language processing are unclear due to the methodology employed. In their experiments, semantic knowledge was acquired from a corpus containing the materials to be segmented in the test phase. One could speculate whether the largest part of the benefit obtained thanks to the addition of semantic knowledge was not due to this hyper-specificity of the LSA corpus (i.e., the inclusion of the test materials). If this were the case, it would call into question the possibility of using LSA to acquire generic semantic knowledge that can be used to segment new texts. A priori, the problem does not seem serious for at least two reasons. First, Choi et al.'s segmentation procedure does not rely on supervised learning in which a system learns how to efficiently segment a text from training data. The LSA corpus only intervenes in an indirect manner by allowing the extraction of semantic proximities between words that are then used to compute similarities between parts of the text to segment (see Section 2 for details). Second, Choi et al. employed a large number of small test samples to evaluate their algorithm, each making up—on average—0.15% of the LSA corpus. The present study shows, however, that the presence of the test materials in the LSA corpus has an important effect, but also that the generic semantic knowledge derived from large corpora clearly improves the segmentation accuracy. This conclusion is drawn from two experiments in which the presence or absence of the test materials in the LSA corpus is manipulated. The first experiment is based on the original materials from Choi et al., which consisted of a small corpus (1,000,000 words). The second experiment is based on a much larger corpus (25,000,000 words). Before reporting these experiments, Choi's algorithm and the use of LSA within this framework are described.

2. The Two Versions of Choi's Algorithm

The segmentation algorithm proposed by Choi (2000) is made up of the three steps usually found in any segmentation procedure based on lexical cohesion. Firstly, the document to be segmented is divided into minimal textual units, usually sentences. Then, a similarity index between every pair of adjacent units is calculated. Each raw similarity value is cast on an ordinal scale by taking the proportion of neighboring values that are smaller than it. Lastly, the document is segmented recursively according to the boundaries between the units that maximize the sum of the average similarities inside the segments thus comprised (divisive clustering).

The step of greatest interest here is the one that calculates the inter-sentence similarities. The procedure initially proposed by Choi (2000), C99, rests exclusively on the information contained in the text to be segmented. According to the vector space model, each sentence is represented by a vector of word frequency count, and the similarity between two sentences is calculated by means of the cosine measure between the corresponding vectors. In a first evaluation based on the procedure described below, Choi showed that its algorithm outperforms several other approaches such as TextTiling (Hearst 1997) and Segmenter (Kan, Klavans, and McKeown 1998).

Choi et al. (2001) claimed that it was possible to improve the inter-sentence similarities index by taking into account the semantic proximities between words estimated on the basis of Latent Semantic Analysis (LSA). Briefly stated, LSA rests on the thesis that

analyzing the contexts in which words occur permits an estimation of their similarity in meaning (Deerwester et al. 1990; Landauer and Dumais 1997). The first step in the analysis is to construct a lexical table containing an information-theoretic weighting of the frequencies of the words occurrence in each document (i.e. sentence, paragraph, or text) included in the corpus. This frequency table undergoes a Singular Value Decomposition that extracts the most important orthogonal dimensions, and, consequently, discards the small sources of variability in term usage. After this step, every word is represented by a vector of weights indicating its strength of association with each of the dimensions. This makes it possible to measure the semantic proximity between any two words by using, for instance, the cosine measure between the corresponding vectors. Proximity between any two sentences (or any other textual units), even if these sentences were not present in the original corpus, can be estimated by computing a vector for each of these units—which corresponds to the weighted sum of the vectors of the words that compose it—and then by computing the cosine between these vectors (Deerwester et al. 1990). Choi et al. (2001) have shown that using this procedure to compute the inter-sentence similarities results in the previous version of the algorithm (based solely on word repetition) being outperformed.

3. Experiment 1

The aim of this experiment is to determine the impact of the presence of the test materials in the LSA corpus on the results obtained by Choi et al. (2001). Does semantic knowledge acquired from a corpus that does not include the test materials also improve the segmentation accuracy?

3.1 Method

This experiment was based on the procedure and test materials designed by Choi (2000), which was also used by several authors as a benchmark for comparing segmentation systems (Brants et al. 2002; Ferret 2002; Kehagias et al. 2003; Utiyama and Isahara 2001). The task consists in finding the boundaries between concatenated texts. Each test sample is a concatenation of ten text segments. Each segment consisted in the first n sentences of a randomly selected text from two sub-sections of the Brown corpus. For the present experiment, I used the most general test materials built by Choi (2000), in which the size of the segments within each sample varies randomly from 3 to 11 sentences. It is composed of 400 samples.

The analysis related to the comparison between the accuracy of the algorithm when the test materials were included in the LSA corpus (*Within*) and when it was not (*Without*). One *Within* semantic space, which corresponds to the one used by Choi et al., was built using the entire Brown corpus as the LSA corpus. Four hundred different *Without* spaces were built, one for each test sample, by each time removing from the Brown corpus only the sentences that make this sample.

To extract the LSA space and to apply the segmentation algorithm, a series of parameters had to be set. First of all, paragraphs were used as documents for building the lexical tables because Choi et al. observed that such middle-sized units were more effective than shorter units (i.e., sentences). The words on Choi's stoplist were removed, as were those that appeared only once in the whole corpus. Words were not stemmed, as in Choi et al. (2001). To build the LSA space, the singular value decomposition was realized using the program SVDPACKC (Berry 1992; Berry et al. 1993), and the first

Table 1
Error rates and variance (in parentheses) for the *Within* and the *Without* conditions.

	Pk	WindowDiff
Within	0.084 (0.005)	0.090 (0.005)
Without	0.120 (0.006)	0.126 (0.006)

300 singular vectors were retained. Concerning the segmentation algorithm, I used the version in which the number of boundaries to be found is imposed, and thus fixed at nine. An 11×11 rank mask was used for the ordinal transformation, as recommended by Choi (2000).

3.2 Results

The segmentation accuracy was evaluated by means of the index reported by Choi et al. (2001): the Pk measure of segmentation inaccuracy (Beeferman, Berger, and Lafferty 1999), which gives the proportion of sentences that are wrongly predicted to belong to the same segment or wrongly predicted to belong to different segments. I also report, for potential future comparison, Pevzner and Hearst's (2002) WindowDiff index, which remedies several problems in the Pk measure.

Results are provided in Table 1.[1] Compared with the *Within* condition, the performance in the *Without* condition is definitely worse, as confirmed by t tests for paired sample (each test sample being used as an observation) that are significant for an alpha smaller than 0.0001. The C99 algorithm, which does not employ LSA to estimate the similarities between the sentences, produces a Pk of 0.13 (Choi et al. 2001, Table 3, line 3: No stemming). It appears that although the *Without* condition is still better than C99, the benefit is very small.

Before concluding that the presence of the test materials in the LSA corpus strongly modified the semantic space, an alternative explanation must be considered. The loss of accuracy in the *Without* condition could potentially be due to the fact that the words indexed in the corresponding LSA spaces are systematically slightly fewer than those present in the *Within* space. Removing each test sample led to the loss—on average—of 23 different words out of the 25,847 words that are indexed in the *Within* space. In the *Without* spaces, these words are no longer available to estimate the similarity of the sentences, whereas they are employed in the *Within* space. In order to determine whether this factor can explain the difference in performance, a complementary analysis was carried out on the *Within* space in which, for each test sample, only the words present in the corresponding *Without* space were taken into account. In this manner, only the semantic relations can come into play. Compared with the complete *Within* space, almost no drop in performance was observed: the Pk error rate went from 0.084 to 0.085 in the new analysis. This result indicates that it is not the words selected for the calculation of the proximities that matter, but the semantic relations in the spaces extracted from the word co-occurrences by the Singular Value Decomposition.

1 The error rate is in fact slightly better than that reported by Choi et al. (2001). The difference could be due to several factors, such as the pre-processing of the Brown corpus (e.g., tokenization and paragraph identification) or the scaling function applied to the raw frequencies, which was here the standard information-theoretic weighting described in Landauer, Foltz, and Laham (1998).

4. Experiment 2

Experiment 1 was conducted on the Choi et al. (2001) LSA corpus, a 1,000,000-word collection of texts from very different genres and with varied themes. The smallness of the corpus and diversity of the texts could have affected the results at two levels. First, removing a few sentences of a text should have less impact if the corpus contains a lot of texts on similar topics. Second, a larger corpus would probably also permit the extraction of a more stable and efficient semantic space. This could produce a greater difference between the LSA version of the algorithm and the version that does not use additional semantic knowledge (C99). For these reasons, a second experiment was conducted on the basis of a much larger corpus consisting of the articles published during 1997 and 1998 in the Belgian French-speaking newspaper Le Soir (roughly 52,000 articles and 26,000,000 words). In this corpus, the test materials from each sample account for—on average—0.0066% of the complete corpus. This second experiment also made it possible to compare the *Within* and *Without* spaces with a *Former* space composed of articles published in the same newspaper, but during the years 1995 and 1996 (roughly 50,000 articles and more than 22,000,000 words). This condition will show the possibility of using LSA to build even more generic semantic knowledge, since the LSA corpus is earlier than the text to segment.

4.1 Method

The test materials were extracted from the 1997–1998 corpus following the guidelines given in Choi (2000). It is composed of 400 samples of ten segments, of which the length varies randomly from 3 to 11 sentences. Three types of LSA space were composed. The *Within* space is based on the whole 1997–1998 corpus. Four hundred different *Without* spaces were built as described in Experiment 1. Finally, a *Former* space was built from the 1995–1996 corpus. The parameters employed to build the semantic spaces are identical to those used in Experiment 1 with one exception: in order to reduce the size of the lexical tables the whole articles and not the paragraphs were used as documents.

4.2 Results

Although the results are mostly similar to those obtained in Experiment 1, Table 2 shows some interesting differences. The discrepancy between the *Within* and *Without* condition is much smaller, even if it remains statistically significant ($p < 0.0001$). Using a corpus from the same source, but with earlier years, still returns a poorer performance ($p < 0.0001$). The C99 algorithm, which is not based on LSA, produces a Pk error rate of 0.150, a value definitely worse than those obtained with the *Without* and *Former* spaces. This confirms the usefulness of semantic knowledge acquired from large corpora in estimating inter-sentence similarities.

5. Conclusion

The two experiments showed that the presence of the test materials in the LSA corpus increases the algorithm accuracy even when a corpus of more than 25,000,000 words is used. They also showed that the use of independent semantic knowledge improves the segmentation accuracy and that this can be observed even when the semantic knowledge is extracted from former years of the same source. This observation underlines

Table 2
Error rates and variance (in parentheses) for the *Within*, *Without* and *Former* conditions.

	Pk	WindowDiff
Within	0.069 (0.004)	0.073 (0.004)
Without	0.080 (0.004)	0.085 (0.005)
Former	0.097 (0.005)	0.101 (0.005)

the possibility of building relatively generic semantic knowledge; that is, knowledge which could be employed to process new linguistic data, as has been recently proposed in a anaphora resolution algorithm, in a continuous speech recognition system, or in machine translation (Bellegarda 2000; Klebanov and Wiemer-Hastings 2002; Kim, Chang, and Zhang 2003). A question the present study does not answer concerns the possibility of employing a corpus drawn from another source, such as another newspaper. Bellegarda (2000) observed in speech recognition tasks that such a semantic space is definitely less effective. It is nevertheless possible that evaluating the semantic proximity between two sentences is less affected by the style of composition of the source than predicting the next word of a statement.

Recently, several authors have proposed segmentation algorithms, based mainly on dynamic programming, that equal or even outperform Choi's results (Ji and Zha 2003, Kehagias et al. 2003; Utiyama and Isahara 2001). These algorithms do not rest on additional semantic knowledge. According to the results of the present study, they could still be improved by taking into account such knowledge.

Finally, this study allows a more general conclusion about the use of LSA for natural language processing. If one's objective is to analyze a linguistic phenomenon in a large corpus such as for instance the factors determining the use of causal connectives (Degand, Spooren, and Bestgen 2004), it is preferable to extract the semantic space from the corpus at hand. The two experiments did indeed show that such specific corpora allow the extraction of a more efficient semantic space. However, if the objective is to test the effectiveness of an algorithm intended to process new linguistic data on the basis of a semantic space built beforehand, one must avoid including the material to analyze in the LSA corpus since that would produce an over-estimate of the effectiveness of the procedure.

Acknowledgments
Yves Bestgen is research fellow of the Belgian National Fund for Scientific Research (FNRS). This work was supported by grant FRFC 2.4535.02 and by a grant (Action de Recherche concerte) of the government of the French-language community of Belgium. A previous version was presented at TALN 2005 (Traitement Automatique des Langues Naturelles, Dourdan, France). Thanks to the anonymous reviewers for their valuable comments.

References
Beeferman, Doug, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210.

Bellegarda, Jerome R. 2000. Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, 8(1):78–84.

Berry, Michael W. 1992. Large scale singular value computations. *International Journal of Supercomputer Applications*, 6(1):13–49.

Berry, Michael W., Theresa Do, Gavin O'Brien, Vijay Krishna, and Sowmini Varadhan. 1993. SVDPACKC: version 1.0 user's guide. Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

Bolshakov, Igor A. and Alexander Gelbukh. 2001. Text segmentation into paragraphs based on local text cohesion. In *Proceedings of Text, Speech and Dialogue (TSD-2001)*. Springer-Verlag, Berlin, pages 158–166.

Brants, Thorsten, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of CIKM'02*, McLean, VA, pages 211–218.

Choi, Freddy Y. Y. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL-00*, Seattle, WA, pages 26–33.

Choi, Freddy Y. Y., Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent semantic analysis for text segmentation. In *Proceedings of NAACL'01*, Pittsburgh, PA, pages 109–117.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Degand, Liesbeth, Wilbert Spooren, and Yves Bestgen. 2004. On the use of automatic tools for large scale semantic analyses of causal connectives. In *Proceedings of ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain, pages 25–32.

Ferret, Olivier. 2002. Using collocations for topic segmentation and link detection. In *Proceedings of COLING 2002*, Taipei, Taiwan, pages 260–266.

Hearst, Marti. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Heinonen, Oskari. 1998. Optimal multi-paragraph text segmentation by dynamic programming. In *Proceedings of 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, Canada, pages 1484–1486.

Ji, Xiang and Hongyuan Zha. 2003. Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, pages 322–329.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear segmentation and segment significance. In *Proceedings of the 6th International Workshop of Very Large Corpora*, Montreal, Canada, pages 197–205.

Kaufmann, Stefan. 1999. Cohesion and collocation: Using context vectors in text segmentation. In *Proceedings of ACL'99*, College Park, MD, pages 591–595.

Kehagias, Athanasios, Fragkou Pavlina, and Vassilios Petridis. 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*, Budapest, Hungary, pages 171–178.

Kim, Yu-Seop, Jeong-Ho Chang, and Byoung-Tak Zhang. 2003. An empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. In *Lecture Notes in Computer Science,* vol. 2637. Springer-Verlag, Berlin, pages 111–116.

Klebanov, Beata and Peter M. Wiemer-Hastings. 2002. Using LSA for pronominal anaphora resolution. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, vol. 2276, Springer-Verlag, Berlin, pages 197–199.

Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, pages 286–288.

Landauer, Thomas K. and Susan T. 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, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. An Introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284.

Lin, Ming, Jay F. Nunamaker, Michael Chau, and Hsinchun Chen. 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 (HICSS-37)*, Big Island, HI, Track 1, Volume 1, pages 10003.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Pevzner, Lev and Marti Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Ponte, Jay M. and W. Bruce Croft. 1997. Text segmentation by topic. In *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries*, *Lecture Notes in Computer Science*, vol. 1324, Springer-Verlag, Berlin, pages 120–129.

Utiyama, Masao and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of ACL'2001*, Toulouse, France, pages 491–498.