Proceedings of HLT-NAACL 2003
Student Research Workshop , pp. 49-54
Edmonton, May-June 2003

Spoken and Written News Story Segmentation using Lexical Chains

Nicola Stokes.
Department of Computer Science,
University College Dublin, Ireland.
Nicola.Stokes@ucd.ie

Abstract

In this paper we describe a novel approach to lexical chain based segmentation of broadcast news stories. Our segmentation system SeLeCT is evaluated with respect to two other lexical cohesion based segmenters TextTiling and C99. Using the P_k and *WindowDiff* evaluation metrics we show that SeLeCT outperforms both systems on spoken news transcripts (CNN) while the C99 algorithm performs best on the written newswire collection (Reuters). We also examine the differences between spoken and written news styles and how these differences can affect segmentation accuracy.

1 Introduction

Text segmentation can be defined as the automatic identification of boundaries between distinct textual units (segments) in a textual document. The aim of early segmentation research was to model the discourse structure of a text, thus focusing on the detection of fine-grained topic shifts, at a clausal, sentence or passage/subtopic level (Hearst 1997). More recently with the introduction of the TDT initiative (Allan et al. 1998) segmentation research has concentrated on the detection of coarse-grained topic shifts resulting in the identification of story boundaries in news feeds. In particular, unsegmented broadcast news streams represent a challenging real-world application for text segmentation approaches, since the success of other tasks such as topic tracking or first story detection depend heavily on the correct identification of distinct and non-overlapping news stories. Most approaches to story segmentation use either Information Extraction techniques (cue phrase extraction), techniques based on lexical cohesion analysis or a combination of both (Reynar 1998; Beeferman et al. 1999). More recently promising results have also been achieved though the use of Hidden Markov modeling techniques, which are commonly used in speech recognition applications (Mulbregt et al. 1999).

In this paper we focus on lexical cohesion based approaches to story segmentation. Lexical cohesion is one element of a broader linguistic device called cohesion which is describe as the textual quality responsible for making the elements of a text appear unified or connected. More specifically, lexical cohesion 'is the cohesion that arises from semantic relationships between words' (Morris, Hirst 1991). With respect to segmentation, an analysis of lexical cohesion can be used to indicate portions of text that represent single topical units or segments i.e. they contain a high number of semantically related words. Almost all approaches to lexical cohesion based segmentation examine patterns of syntactic repetition in the text e.g. (Reynar 1998; Hearst 1997; Choi 2000). However, there are four additional types of lexical cohesion present in text: synonymy (*car, automobile*), specialization/generalization (*horse, stallion*), part-whole/whole-part (*politicians, government*) and statistical co-occurrences (*Osama bin Laden, World Trade Center*). Lexical chaining based approaches to text segmentation, on the other hand, analyse all aspects of lexical cohesion in text. Lexical chains are defined as groups of semantically related words that represent the lexical cohesive structure of a text e.g. {*flower, petal, rose, garden, tree*}. In our lexical chaining implementation, words are clustered based on the existence of statistical relationships and lexicographical associations (provided by the WordNet online thesaurus) between terms in a text.

There have been three previous attempts to tackle text segmentation using lexical chains. The first by Okumara and Honda (1994) involved an evaluation based on five Japanese texts, the second by Stairmand (1997) used twelve general interest magazine articles and the third by Kan et al. (1998) used fifteen *Wall Street Journal* and five *Economist* articles. All of these attempts focus on sub-topic rather than story segmentation. In contrast, this paper investigates the usefulness of lexical chains as a technique for determining story segments in spoken and written broadcast news streams. In Section 2, we explain how this technique can be refined

to address story segmentation. In Section 3, we compare the segmentation performance of our lexical chaining algorithm with two other well known lexical cohesion based approaches to segmentation; namely TextTiling (Hearst 1997) and C99 (Choi 2000). Finally we examine the grammatical differences between written and spoken news media and show how these differences can be utilized to improve spoken transcript segmentation accuracy.

2 SeLeCT: Segmentation using Lexical Chains on Text

In this section we present our topic segmenter SeLeCT. This system takes a concatenated stream of text and returns a segmented stream of distinct news reports. The system consists of three components a 'Tokeniser', a 'Chainer' which creates lexical chains, and a 'Detector' that uses these chains to determine news story boundaries. More detailed descriptions of the 'Tokeniser' and 'Chainer' components are reported in Stokes et al. (2003).

2.1 The Tokeniser

The objective of the chain formation process is to build a set of lexical chains that capture the cohesive structure of the input stream. Before work can begin on lexical chain identification, each sample text is processed by a part-of-speech tagger. Morphological analysis is then performed on these tagged texts; all plural nouns are transformed into their singular form, adjectives pertaining to nouns are nominalized and all sequences of words that match grammatical structures of compound noun phrases are extracted. This idea is based on a simple heuristic proposed by Justeson and Katz (Justeson, Katz 1995), which involves scanning part-of-speech tagged texts for patterns of adjacent tags that commonly match proper noun phrases like 'White House aid', 'PLO leader Yasir Arafat', and WordNet noun phrases like 'red wine' or 'act of god'. Since the likelihood of finding exact syntactic matches of these phrases elsewhere in a story is low, we include a fuzzy string matching function in the lexical chainer to identify related phrases like George_*Bush* ⇔ President_*Bush.*

2.2 The Lexical Chainer

The aim of the Chainer is to find relationships between tokens (nouns, proper nouns, compound nouns, nominalized adjectives) in the data set using the WordNet thesaurus and a set of statistical word associations, and to then create lexical chains from these relationships with respect to a set of chain membership rules. The chaining procedure is based on a single-pass clustering algorithm, where the first token in the input stream becomes the head of the first lexical chain. Each subsequent token is then added to the most recently updated chain that it shares the strongest semantic relationship[1] with. This process is continued until all tokens in the text have been chained. Our chaining algorithm is similar to one proposed by St Onge (1995) for the detection of malapropisms in text, however statistical word associations and proper nouns were not considered in his original implementation.

2.3 Boundary Detection

The final step in the segmentation process is to partition the text into its individual news stories based on the patterns of lexical cohesion identified by the Chainer in the previous step. Our boundary detection algorithm is a variation on one devised by Okumara and Honda (Okumara, Honda 1994) and is based on the following observation:

'Since lexical chain spans (i.e. start and end points) represent semantically related units in a text, a high concentration of chain begin and end points between two adjacent textual units is a good indication of a boundary point between two distinct news stories'

We define boundary strength $w(n, n+1)$ between each pair of adjacent textual unit in our test set, as the sum of the number of lexical chains whose span ends at paragraph n and the number of chains that begin their span at paragraph $n+1$. When all boundary strengths between adjacent paragraphs have been calculated we then get the mean of all the non-zero cohesive strength scores. This mean value then acts as the minimum allowable boundary strength that must be exceeded if the end of textual unit n is to be classified as the boundary point between two news stories.

Finally these boundary strength scores are 'cleaned' using an error reduction filter which removes all boundary points which are separated by less than x number of textual units from a higher scoring boundary, where x is too small to be a 'reasonable' story length. This filter has the effect of smoothing out local maxima in the boundary score distribution, thus increasing segmentation precision. Different occurrences of this error are illustrated in Figure 1. Regions A and C represent clusters of adjacent boundary points. In this situation only the boundary with the highest score in the cluster is retained as the true story boundary. Therefore the boundary which scores 6 is retained in region A while in region C both points have the same score so in this case we consider the last point in region C to be the correct boundary position. Finally, the story boundary in region B is also eliminated because it is situated too close to the boundary points in

[1] Repetition is the strongest cohesive relationship, followed by synonymy, and then statistical associations, generalization/specialization and part-whole/whole-part relationships.

region C and it has a lower score than either of those boundaries.

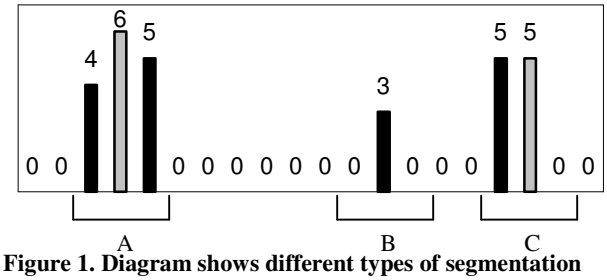

Figure 1. Diagram shows different types of segmentation error; numbers greater than zero are possible boundary positions, while zero scores represent no story boundary point between these two textual units.

3 Segmentation Evaluation

In this section we give details of two news story segmentation test sets, some evaluation metrics used to determine segmentation accuracy, and the performance results of the SeLeCT, C99 and TextTiling algorithms.

3.1 News Segmentation Test Collections

Both the CNN and Reuters test collections referred to in this paper contain 1000 randomly selected news stories taken from the TDT1 corpus. These test collections were then reorganized into 40 files each consisting of 25 concatenated news stories. Consequently, all experimental results in Section 3.3 are averaged scores generated from the individual results calculated for each of the 40 samples. By definition a segment in this context refers to a distinct news story, thus eliminating the need for a set of human-judged topic shifts for assessing system accuracy.

3.2 Evaluation Metrics

There has been much debate in the segmentation literature regarding appropriate evaluation metrics for estimating segmentation accuracy. Earlier experiments favored an IR style evaluation that measures performance in terms of recall and precision. However these metrics were deemed insufficiently sensitive when trying to determine system parameters that yield optimal performance. The most widely used evaluation metric is Beeferman et al.'s (1999) **probabilistic error metric P_k,** which calculates segmentation accuracy with respect to three different types of segmentation error: false positives (falsely detected segments), false negatives (missed segments) and near-misses (very close but not exact boundaries). However, in a recent publication Pevzner and Hearst (2002) highlight several faults with the P_k metric. Most notable they criticize P_k for its unfair penalization of false negatives over false positives and its over-penalization of near-misses. In their paper, the authors proposed an alternative error metric called ***WindowDiff*** which rectifies these problems.

3.3 Story Segmentation Results

In this section we present performance results for each segmenter on both the CNN and Reuters test sets with respect to the aforementioned evaluation metrics. As explained in Section 3, we determine the effectiveness of our SeLeCT system with respect to two other lexical cohesion based approaches to segmentation, namely the TextTiling (Hearst 1997) and C99 algorithms (Choi 2000)[2]. We also include average results from a random segmenter that returned 25 random boundary positions for each of the 40 files in both test sets. These results represent a lower bound on segmentation performance. All results in this section are calculated using paragraphs as the basic unit of text. Since both our test sets are in SGML format, we consider the beginning of a paragraph in this context to be indicated by a speaker change tag in the CNN transcripts and a paragraph tag in the case of the Reuters news stories.

System	CNN		Reuters	
	P_k	WD	P_k	WD
SeLeCT	***0.25***	***0.253***	0.191	0.207
TextTiling	0.259	0.299	0.221	0.244
C99	0.294	0.351	***0.128***	***0.148***
Random	0.421	0.48	0.490	0.514

Table 1: P_k and WD (WindowDiff) values for segmentation systems on CNN and Reuters Collections.

Table 1 summarizes the results of the CNN data set for each segmentation system evaluated with respect to the four metrics. All values for these metrics range from 0 to 1 inclusively, where 0 represents the lowest possible measure of system error. From these results we observe that the accuracy of our SeLeCT segmentation algorithm is greater than the accuracy of C99, TextTiling or the Random segmenter for both evaluation metrics on the CNN 'spoken' data set. As for the Reuters segmentation performance, the C99 algorithm significantly outperforms both the SeLeCT and TextTiling systems. We also observe that the *WindowDiff* metric penalizes systems more than P_k, however the overall ranking of the systems with respect to these error metrics remains the same. With regard to the SeLeCT system, optimal performance was achieved when only patterns of lexical repetition were examined during the boundary detection phase, thus eliminating the need for an examination of lexicographical and statistical relationships between tokens in the text.

[2] We use Choi's java implementations of TextTiling and C99 available for free download at www.cs.man.ac.uk/~choif. In (Choi 2000) boundaries are hypothesized using sentences as the basic unit of text; however both C99 and TextTiling can take advantage of paragraph information when the input consists of one paragraph per line.

A similar conclusion was reported by Hearst (1997) and Min-Yen et al. (1998); however neither of these approaches included statistical word associations in their chaining process.

4 Written and Spoken Text Segmentation

It is evident from the results of our segmentation experiments on the CNN and Reuters test collections that system performance is dependant on the type of news source being segmented i.e. spoken texts are more difficult to segment. This disagreement between result sets is a largely unsurprising outcome as it is well documented by the linguistic community that written and spoken language modes differ greatly in the way in which they convey information. At a first glance, it is obvious that written texts tend to use more formal and verbose language than their spoken equivalents. However, although CNN transcripts share certain spoken text characteristics (see Section 4.1), they lie somewhere nearer written documents on a spectrum of linguistic forms of expression, since they contain a mixture of speech styles ranging from formal prepared speeches from anchor people, politicians, and correspondents, to informal interviews/comments from ordinary members of the public. Furthermore, spoken language is also characterized by false starts, hesitations, back-trackings, and interjections; however information regarding prosodic features and these characteristics are not represented in CNN transcripts. In the next section we look at some grammatical differences between spoken and written text that are actually evident in CNN transcripts. In particular, we look at the effect that these differences have on parts of speech distributions and how these impact segmentation performance.

4.1 Lexical Density

One method of measuring the grammatical intricacy of speech compared to written text, is to calculate the lexical density of the language being used. The simplest measure of lexical density, as defined by Halliday (1995), is the 'the number of lexical items (content words) as a portion of the number of running words (grammatical words)'. Halliday states that written texts are more lexically dense while spoken texts are more lexically sparse. In accordance with this, we observe based on part-of-speech tag information that the CNN test set contains 8.58% less lexical items than the Reuters news collection.[3]

[3] Lexical items included all nouns, adjectives and verbs, except for function verbs like modals and auxiliary verbs. Instead these verbs form part of the grammatical item lexicon with all remaining parts of speech. Our CNN and Reuters data sets consisted of 43.68% and 52.26% lexical items respectively.

Halliday explains that this difference in lexical density between the two modes of expression can be attributed to the following observation:
'Written language represents phenomena as products, while spoken language represents phenomena as processes.'
In real terms this means that written text tends to conveys most of its meaning though nouns (NN) and adjectives (ADJ), while spoken text conveys it though adverbs (ADV) and verbs (VB). To illustrate this point consider the following written and spoken paraphrase of the same information:

Written: **Improvements/NN** in American zoos have resulted in better **living/ADJ** conditions for their **animal residents/NN.**

Spoken: **Since/RB** American zoos have been **improved/VB** the animals **residing/VB** in them are **now/RB living/VB** in better conditions.

Although this example is a little contrived, it shows that in spite of changes to the grammar, by and large the vocabulary has remained the same. More specifically, these paraphrases illustrate how the products in the written version*, improvements, resident,* and *living*, are conveyed as processes in spoken language though the use of verbs. The spoken variant also contains more adverbs; a grammatical necessity that provides cohesion to text when processes are being described in verb clauses.

As explained in Section 2.2 the SeLeCT lexical chainer only looks at cohesive relationships between nouns and nominalized adjectives in a text. This accounts partly for SeLeCT's lower performance on the CNN test set, since the extra information conveyed though verbs in spoken texts is ignored by the lexical chainer. However since C99 and TextTiling use all parts of speech in their analysis of the text, the replacement of products with processes is not the reason for a similar deterioration in their performance. More specifically, both C99 and TextTiling rely on stopword lists to identifying spurious inter-segment links between function words that by their nature do not indicate common topicality. For the purpose of their original implementation their stopwords lists contained mostly pronouns, determiners, adverbs, and function verbs such as auxiliary and modal verbs. However, we have observed that the standard set of textual function verbs is not enough for speech text processing tasks and that their lists should be extended to include other common 'low information' verbs. These types of verbs are not necessarily characterized by large frequency counts in the spoken news collection like the domain specific phrases *to report* or *to comment*. Instead these verbs tend to have no 'equivalent' nominal form, like the verbs '*to let' 'to hear' 'to look'* or *'to try'*.

To test this observation we re-ran C99 and TextTiling experiments on the Reuters and CNN

collections, using only nouns, adjectives, nominalized verbs (provided by the NOMLEX (Meyers et al. 1998)), and nominalized adjectives as input. Our results show that there is a significant decrease in *WindowDiff* error for the C99 system on both the CNN collection (a decrease from 0.351 to 0.268) and the Reuters collection (a decrease from 0.148 to 0.121). Similarly, we observe an improvement in the *WindowDiff* based performance of the TextTiling system on the CNN data set (a decrease from 0.299 to 0.274). However, we observe a marginal fall in performance on the Reuters data set (an increase from 0.244 to 0.247). These results illustrate the increased dominance of verbs in spoken text and the importance of function verb removal by our verb nominalization process for CNN segmentation performance.

4.2 Reference and Conjunction in Spoken Text

A picture paints a thousand words, they say, and since news programme transcripts are accompanied by visual and audio cues in the news stream, there will always be a loss in communicative value when transcripts are interpreted independently. As stated in Section 4.1, it is well known that conversational speech is accompanied by prosodic and paralinguistic contributions, facial expressions, gestures, intonation etc., which are rarely conveyed in spoken transcripts. However there are also explicit (exophoric) references in the transcript to events occurring outside the lexical system itself. These exophoric references in CNN transcripts relate specifically to audio references like speaker change, musical interludes, background noise; and visual references like event, location and people shots in the video stream. We believe that this property of transcribed news is another reason for the deterioration in segmentation performance on the CNN test collection.

Solving endophoric (anaphora and cataphora) and exophoric reference has long been recognized as a very difficult problem, which requires pragmatic, semantic and syntactic knowledge in order to be solved. However there are simple heuristics commonly used by text segmentation algorithms that in our case can be used to take advantage of the increased presence of reference in spoken text. One such heuristic is based on the observation that when common referents like personal and possessive pronouns, and possessive determiners appear at the beginning of a sentence, this indicates that these referents are linked in some way to the previous textual unit (in our case the previous paragraph). The resolution of these references is not of interest to our algorithm but the fact that two textual units are linked in this way gives the boundary detection process an added advantage when determining story segments in the text. An analysis of conjunction (another form of textual cohesion) can also be used to provide the detection process with useful evidence of related paragraphs, since paragraphs that begin with conjunctions (*because*, *and*, *or*, *however*, *nevertheless*) and conjunctive phrases (*in the mean time*, *in addition*, *on the other hand*) are particularly useful in identify cohesive links between units in conversational/interview sequences in the transcript.

4.3 Refining SeLeCT Boundary Detection

In Section 2.3 we describe in detail how the boundary detection phrase uses lexical chaining information to determine story segments in a text. One approach to integrating referential and conjunctive information with the lexical cohesion analysis provided by the chains is to remove all paragraphs from the system output that contain a reference or conjunctive relationship with the paragraph immediately following it in the text. The problem with this approach is that P_k and *WindowDiff* errors will increase if 'incorrect' segment end points are removed that represented near system misses rather than 'pure' false positives. Hence, we take a more measured approach to integration that uses conjunctive and referential evidence in the final filtering step of the detection phrase, to eliminate boundaries in boundary clusters (Section 2.3) that cannot be story end points in the news stream. Figure 2 illustrates how this technique can be used to refine the filtering step. Originally, the boundary with score six in region A would have been considered the correct boundary point. However since a conjunctive phrase links the adjacent paragraphs at this boundary position in the text, the boundary which scores five is deemed the correct boundary point by the algorithm.

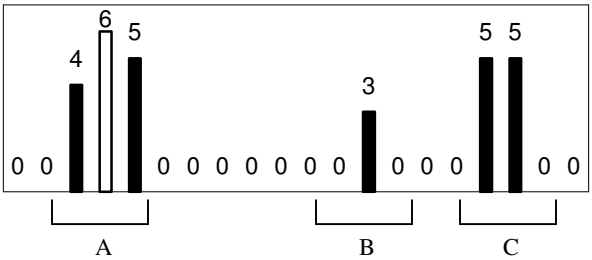

Figure 2 Illustrates how cohesion information can help SeLeCT's boundary detector resolve clusters of possible story boundaries.

Using this technique and the verb nominalization process described in section 4.1 on both news media collections, we observed an improvement in SeLeCT system performance on the CNN data set (a decrease in error from 0.253 to 0.225), but no such improvement on the Reuters collection. Again the ineffectiveness of this technique on the Reuters results can be attributed to differences between the two modes of language expression, where conjunctive and referential relationships resolve 51.66% of the total possible set of boundary points between stories in the CNN collection and only 22.04% in the Reuters collection. In addition, these references in the Reuters articles mostly occur between sentences in a paragraph rather than between paragraphs in the text thus provide no additional cohesive

information. A summary of the improved results discussed in this section is shown in Table 2.

System	CNN WD Score		Reuters WD Score	
	Before	After	Before	After
SeLeCT	0.253	0.225	0.207	0.209
C99	0.351	0.268	0.148	0.121
TextTiling	0.299	0.274	0.244	0.247

Table 2: Improvements in system performance as a result of system modifications discuss in Sections 4.1 and 4.3.

5 Conclusions

In this paper we have presented a lexical chaining based approach to coarse-grained segmentation of CNN news transcripts and concatenated Reuters newswire articles. We have shown that the performance of our SeLeCT system exceeds that of the TextTiling and C99 systems when detecting topic shifts in CNN transcripts. However the results of a similar experiment on Reuters news stories showed that the C99 system outperformed all other systems on a written news collection. Overall, lower CNN segmentation results were attributed to the information loss caused by prosodic and paralinguistic characteristics of speech and grammatical differences between written and spoken modes of expression. Further experiments showed that by limiting the input of all the segmentation systems to nouns, adjectives, and nominalized verbs and adjectives, the effect of these grammatical differences on CNN segmentation performance was significantly reduced. Additional SeLeCT performance improvements were also achieved by using referential and conjunctive relationships as additional evidence of cohesion in the boundary detection step. In future experiments we plan to compare SeLeCT's performance on written and spoken news texts with two recently proposed systems, U00 (Utiyama 2001) and CWM (Choi 2001), which have marginally outperformed the C99 algorithm on Choi's (2000) test corpus.

Acknowledgements

The support of Enterprise Ireland is gratefully acknowledged. Also I wish to thank Marti Hearst for providing us with a version of the *WindowDiff* evaluation software and Joe Carthy for invaluable comments.

References

Allan J., J. Carbonell, G. Doddington, J. Yamron, Y. Yang. *Topic Detection and Tracking Pilot Study Final Report.* In the proceedings of the DARPA Broadcasting News Workshop, pp. 194-218, 1998.

Beeferman D., A. Berger, and J. Lafferty. *Statistical models for text segmentation. Machine Learning*, (34):177-210. 1999.

Choi F., *Advances in domain independent linear text segmentation*. In Proceedings of NAACL'00. 2000.

Choi F., P. Wiemer-Hastings, J. Moore. Latent semantic analysis for Text Segmentation. In proceedings EMNLP 2001, pp.109-117, 2001.

Halliday M.A.K., Spoken and Written Language.Oxford University Press, 1985.

Hearst M., TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, 23 (1):33-64, 1997.

Justeson, J. S., S.M. Katz., Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering (11): 9-27, 1995.

Kan Min-Yen, J. L. Klavans, K. R. McKeown. *Linear Segmentation and Segment Relevance*. In the proceedings of WVLC-6, pp. 197-205, 1998.

Kozima H., *Text segmentation based on similarity between words*. In Proceedings of ACL-93, pp. 286-288, 1993.

Meyers A., et al. *Using NOMLEX to produce nominalization patterns for information extraction*. In Proceedings of the COLING-ACL Workshop on Computational Treatment of Nominals, 1998.

Morris J., G. Hirst, *Lexical Cohesion by Thesaural Relations as an Indicator of the Structure of Text*, Computational Linguistics 17(1), 1991.

Okumura M., T. Honda, Word sense disambiguation and text segmentation based on lexical cohesion. In proceedings of COLING-94, pp. 755-761, 1994.

Pevzner, L., and M. Hearst, *A Critique and Improvement of an Evaluation Metric for Text Segmentation*, Computational Linguistics, 28 (1):19-36, 2002.

Reynar J., *Topic Segmentation: Algorithms and Applications*, Ph.D. thesis, Dept. Computer and Information Science, UPenn, 1998.

Stairmand M.A, A Computational Analysis of Lexical Cohesion with Applications in IR, PhD Thesis, Dept. of Language Engineering, UMIST. 1996.

St-Onge D., *Detecting and Correcting Malapropisms with Lexical Chains*, Dept. of Computer Science, University of Toronto, M.Sc. Thesis, 1995.

Stokes N., J. Carthy, A.F. Smeaton. *SeLeCT: A Lexical Cohesion Based News Story Segmentation System.* Technical Report CS02-03, Dept. of Computer Science, University College Dublin, 2003.

Utiyama M., H. Isahara. *A statistical model for domain-independent text segmentation*. In proceedings of ACL-2001, pp.491-498, 2001.

van Mulbregt P., I. Carp, L. Gillick, S. A. Lowe, J. P. Yamron. *Segmentation of Automatically Transcribed Broadcast News Text*, In Proceedings of the DARPA Broadcast News Workshop, 1999.