# Segmenting Broadcast News Streams using Lexical Chains

Nicola Stokes, Joe Carthy,
*Department of Computer Science,*
*University College Dublin,*
*Ireland.*
{nicola.stokes, joe.carthy}@ucd.ie

Alan F. Smeaton,
*School for Computer Applications and*
*Centre for Digital Video Processing,*
*Dublin City University, Ireland.*
asmeaton@computing.dcu.ie

**Abstract**. In this paper we propose a course-grained NLP approach to text segmentation based on the analysis of lexical cohesion within text. Most work in this area has focused on the discovery of textual units that discuss subtopic structure within documents. In contrast our segmentation task requires the discovery of topical units of text i.e. distinct news stories from broadcast news programmes. Our system SeLeCT first builds a set of lexical chains, in order to model the discourse structure of the text. A boundary detector is then used to search for breaking points in this structure indicated by patterns of cohesive strength and weakness within the text. We evaluate this technique on a test set of concatenated CNN news story transcripts and compare it with an established statistical approach to segmentation called TextTiling.

## 1. Introduction

Text segmentation can be defined as the automatic identification of boundaries between distinct textual units (segments) in a textual document. The importance and relevance of this task should not be underestimated, as good structural organisation of text is often a prerequisite to many important tasks that deal with the management and presentation of data. Consider the usefulness of text segments when responding to a user query in an information retrieval task, where users are given short pieces of relevant text rather than vast quantities of semi relevant documents [1]. Summarisation is another task that can be greatly improved by well-segmented text since the aim of this task is to identify pertinent subtopics in a document and then generate a summary, which encapsulates all of these subtopics [2]. The main motivation of our research is to investigate whether our lexical chaining technique can be used to segment television news shows into distinct new stories. Lexical chaining is a linguistic technique that uses an auxiliary resource (in our case the WordNet online thesaurus [3]) to cluster words into sets of semantically related concepts e.g. {*motorbike, car, lorry, vehicle*}. In this paper we endeavour to explain how such constructs can be used to detect topic shifts in CNN broadcast news programmes extracted from the TDT 1 corpus [4]. The research and results discussed here are a preliminary investigation into the suitability of our lexical chaining technique to the detection of course-grained topic shifts resulting in the identification of news story boundaries. We define a topic shift in this context as the boundary point between two distinct topically cohesive stories. Subtopic or more finely grained topic shifts are those that indicate more subtle thematic changes within a news story e.g. A story on Northern Ireland might report on two separate incidents of violence, however our system must identify the relationship between these consecutive subtopics and threat them as a single topical unit by returning only one story segment on Northern Ireland. The end goal is to develop a robust segmenter, which will eventually be integrated with a video segmenter (i.e. a system that segments news programmes based on colour analysis) to facilitate the retrieval and playback of individual news stories in response to user requests in the DCU Físchlár system [5].

In the next section we discuss in more detail the text segmentation problem and some techniques that have been proposed to solve it. We then describe our model for segmentation based on lexical cohesion analysis. Finally we detail our evaluation methodology followed by results and comparisons with another well established approach to exploratory text segmentation called TextTiling [6].

## 2. Text Segmentation

Text segmentation techniques can be roughly separated into two different approaches; those that rely on lexical cohesion and those that rely on statistical information extraction techniques such as cue information extraction (IE) [7]. For IE techniques to work some explicit structure must be present in the text. Manning's segmenter [7] was required to identify boundaries between real estate classified advertisements, which in general will contain the same types of information *'house price'* or *'location'* etc. As Reynar [8] remarks in his segmentation work on the TDT 1 corpus, using domain cues in news transcripts such as *'Good Morning'*, *'stay with us'*, *'welcome back'* or *'reporting from PLACE'* are reliable indicators of topic shifts. However the problem with these domain cues is that they are not only genre-specific conventions used in news transcripts but they are also programme specific as well. For example in Irish news broadcasts in contrast to their American counterparts, news programmes are never *'brought to you by a PRODUCT NAME'*. Newscaster styles also change across news stations, as certain catch phrases are favoured by some individuals more than others. The consequence of this is that new lists of cues must be generated either manually or automatically in which case an annotated corpus is needed. However as Reynar [8] points out significant gains can be achieved by combining cue information with other feature information such as named entities (President Bush, George W. Bush Jr), character n-grams (sequences of word forms of length n), and semantic similarity. Reynar like Beeferman [9] developed a machine learning approach which combines cues in a probabilistic framework. These combination approaches allow the segmentation system to learn the best indicators of segment boundaries from a training set and also how best to combine these features in a theoretical sound framework. An alternative to these statistical approaches is discussed in the following sections.

## 3. Lexical Cohesion

Lexical cohesion is one element of a broader linguistic device called cohesion, which is the textual quality responsible for making the sentences of a text seem 'to hang together' [10]. Here are a number of different forms of lexical cohesion followed by examples from CNN news transcript.

- **Repetition** – Occurs when a word form is repeated again in a later section of the text e.g. "In *Gaza*, though, whether the Middle East's old violent cycles continue or not, nothing will ever look quite the same once Yasir Arafat come to town. We expect him here in the *Gaza Strip* in about an hour and a half, crossing over from Egypt".
- **Repetition through synonymy** – Occurs when words share the same meaning but have two unique syntactical forms. "Four years ago, it passed a domestic violence act allowing *police*, not just the victims, to press charges if they believe a domestic beating took place. In the past, *officers* were frustrated, because they'd arrive on the scene of a domestic fight, there'd be a clearly battered victim and yet, frequently, there'd be no one to file charges."

- **Word association through specialisation/generalisation** – Occurs when a specialised/generalised form of an earlier word is used. "They've put a possible murder weapon in O.J. Simpson's hands; that's something that no one knew before. And it shows that he bought that knife more than a month or two ahead of time and you might, therefore, start the theory of premeditation and deliberation."
- **Word association through part-whole/whole-part relationships** - Occurs when a part-whole/whole-part relationship exists between two words e.g. '*committee*' is made up of smaller parts called '*members*'. "The Senate Finance *Committee* has just convened. *Members* had been meeting behind closed doors throughout the morning and early afternoon."
- **Statistical associations between words** – These types of relationships occur when the nature of the association between two words cannot be defined in terms of the above relationship types. These relationships are most commonly found by word co-occurrence statistics e.g. *Osama bin Laden* and *the World Trade Centre*.

## 4. Lexical Cohesion and Text Segmentation

Research has shown that lexical cohesion is a useful device in the detection of subtopic shifts in texts [6, 11-14]. Its suitability to segmentation is based on the fact that portions of text that contain high numbers of semantically related words (cohesively strong links) generally constitute a single topical unit. So in terms of segmentation, areas of low cohesive strength within a text are good indicators of topic transitions.

Most approaches to segmentation using lexical cohesion rely on only one form of lexical cohesion i.e. repetition. One such system was developed by Hearst [6] called TextTiling. Hearst's algorithm begins by artificially separating text into groups of fixed blocks of pseudo-sentences (also of fixed length). The algorithm uses the cosine similarity[1] metric to measure cohesive strength between adjacent blocks. Depth scores are then calculated for each block based on the similarity scores between a block and those blocks neighbouring it in the text. The algorithm then deduces boundary points from these scores by hypothesising that high depth scores (major drops in similarity) indicate topic boundary points. Another interesting approach that implicitly considers all of the above lexical cohesive types is Ponte and Crofts segmenter [11]. Their segmentation system uses a word co-occurrence technique called LCA (Local Context Analysis) to determine the similarity between adjacent sentences. LCA expands the context surrounding each sentence by finding other words and phrases that occur frequently with these sentence words in the corpus. The authors show that segmentation based on LCA is particularly suited to texts containing extremely short segments which share very few terms due to their brevity. For example, they evaluated their approach on news summaries which had an average sentence length of 2.8. Kaufmann's [12] VecTiling system augments the basic TextTiling algorithm with a more sophisticated approach to determining block similarity, which is closely related to Ponte and Crofts word expansion technique. However instead of LCA, VecTile uses Schutze's WordSpace model [15] to replace words by vectors containing information about the types of contexts that they are most commonly found in.

[1] The cosine similarity is often used in Information Retrieval to find the similarity between documents by measuring the cosine of the angle between two document vectors of term weights derived from the frequency of occurrence of the terms contained in each document.

Lexical chaining on the other hand explicitly considers the first four types of word associations mentioned in Section 3. We use WordNet as our lexical resource to facilitate chain creation. As already mentioned lexical chains are essentially groups of words that were cluster together due to the existence of lexicographical relationships between themselves and at least one other member of the chain. For example in a document concerning cars a typical chain might consist of the following words {*BMW, vehicle, engine, wheel, car, automobile, tire*}, where each word in the chain is directly or indirectly related to another word by a semantic relationship such as *synonymy* (car and automobile are semantically equivalent), *holonymy* (car has-part engine), *hyponymy* (BMW is a specialisation of a car), *meronymy* (tire is part-of a wheel) and *hyponymy* (vehicle is a generalisation of a car). All these associations can be found in the WordNet taxonomy.

Text segmentation is not a novel application for lexical chaining, in fact in their seminal paper on lexical chain creation using thesaural relations from Roget's thesaurus, Morris and Hirst [10] detail an algorithm capable of determining 'subtopic flow by recording where in the discourse the bulk of one set of chains ends and a new set of chains begin'. However it was not until Okumara and Honda's work on the summarisation of Japanese text, that Morris and Hirst's approach to segmentation was implemented [16]. Segmentation research using chains was also briefly discussed by Stairmand in his analysis of lexical cohesion in IR applications [17]. The novel aspect of our research regards the development of a courser-grained segmenter, that ignores subtle subtopic shifts and fragments news programmes into their constitute news stories rather than the subtopics that constitute them.

## 5. **SeLeCT** - **Se**gmentation using **Le**xical **C**haining on **T**ext.

In this section we present our topic segmenter, SeLeCT. This system takes a concatenated stream of news programs and returns segments consisting of single news reports. The system consists of three components a 'Tokeniser', a 'Chainer' which creates lexical chains, and a 'Detector' that uses these chains to find news story boundaries.

### *5.1 The Tokeniser*

The objective of the chain formation process is to build a set of lexical chains that captures the cohesive structure of the input stream. Before work can begin on lexical chain identification each document is processed by a part of speech tagger. Once the nouns in the text have been identified, morphological analysis is performed on these nouns (i.e. all plurals are transformed into their singular state and any compound nouns (consisting of two/three adjacent nouns) are searched for in WordNet). These part of speech tags also identify potential proper noun entries in the WordNet thesaurus. In general news story proper noun phrases will not be present in WordNet, since keeping an up to date repository of such words is a substantial and never ending problem. However phrases such as 'John Glenn' are present and linked to useful terms like 'senator' and 'astronaut' which when used can significantly improve chaining accuracy. Any remaining proper nouns are still useful to the chaining process as they provide a further means of capturing the repetition element of lexical cohesion mentioned in section 3. The resultant tokenised text consisting of nouns and proper noun phrases (including information regarding their location within the text) is then used as input to the chaining phase of the segmentation process.

### *5.2 The Lexical Chainer*

The aim of the Chainer is to find relationships between tokens (nouns, proper nouns, compound nouns) in the data set using the WordNet thesaurus and to then create lexical chains from these associations 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 forms the first lexical chain and each subsequent token is then added to an existing chain if it is related to at least one other token in that chain by any of the lexicographical relationships defined in Section 3. A stronger criterion than simple semantic similarity is imposed on the addition of a token to a chain, where a token must be added to the most recently updated (semantically related) chain.

In addition a relationship between two tokens is only considered valid if it adheres to the following rules:

1. For repetition or synonymy relationships two tokens must appear no more than 600 words away from each other in the original text for the relationship to hold.
2. For all other relationships this distance constraint is set to a maximum of 500 words between tokens since they are weaker associations than repetition.
3. Further constraints are imposed on relationships between words that have a path length greater than 1 in the WordNet taxonomy e.g. *military action* is related to *germ warfare* by the following relationships: *military action* is a generalisation of *war,* which is a generalisation of *bacterial warfare,* which is a generalisation of *germ warfare*.

Note: The word distance thresholds above were empirically chosen so as to yield optimal system results.

Imposing a word distance threshold depending on the association between two related words is important for two reasons. Firstly these thresholds lessen the effect of spurious chains, which are weakly cohesive chains containing misidentified word associations due to the ambiguous nature of the word forms i.e. associating *bank* with *money* when *bank* refers to a *river bank* is an example of misidentification. The creation of these sorts of chains is undesirable as they add noise to the detection of boundaries described in the next section. Secondly due to the temporal nature of news streams, stories related to important breaking-news topics will tend to occur in close proximity in time. If unlimited distance were allowed, even between strongly related words (i.e. where a repetition relationship exists), some chains would span the entire text if two stories discussing the same topic were situated at the beginning and end of a news programme.

In summary our chaining algorithm proceeds as follows, if an 'acceptable' relationship exists between a token and a chain then the token is added to that chain otherwise the token will become the seed of a new chain. This process is continued until all keywords in the text have been chained.

### *5.3 Boundary Detection*

The final step in the segmentation process is to pass all chain information to the boundary detector. Our boundary detection algorithm is a variation on one devised by Okumara and Honda [16] and is based on the following hypothesis:

*'A high concentration of chain begin and end points exist on the boundary between two distinct news stories.'*

We define boundary strength *w(n, n+1)* between each sentence in a text (defined as a unit of text that begins with a capital letter and ends with a full stop), as the product [2] of the number of lexical chains whose span ends at sentence *n* and the number of chains that begin their span at sentence *n+1*. To illustrate how boundary strengths based on lexical cohesion are calculated consider the following piece of text containing one topic shift (all nouns are highlighted), accompanied by lexical chains derived from this text fragment where chain format is:
{word……… | Sentence number: chain start, chain end}

“Coming up tomorrow when the hearing resumes, we hear testimony from the limousine driver that brought O.J. Simpson to the airport- who brought O.J. Simpson to the airport June 12th, the night of the murders. The **president** of Mothers Against Drunk Driving discusses her **organization's** support of sobriety **checkpoints** over the **holiday weekend**. She hopes **checkpoints** will be used all the **time** to limit the **number** of **fatalities** on the **road**.”

{hearing, testimony | 1, 1} {tomorrow, night, **holiday**, **weekend**, **time** | 1, 3}
{airport | 1, 1} {**president**, **organization**| 2, 2} {**checkpoints** | 2, 3} {murders, **fatalities** | 1, 3}

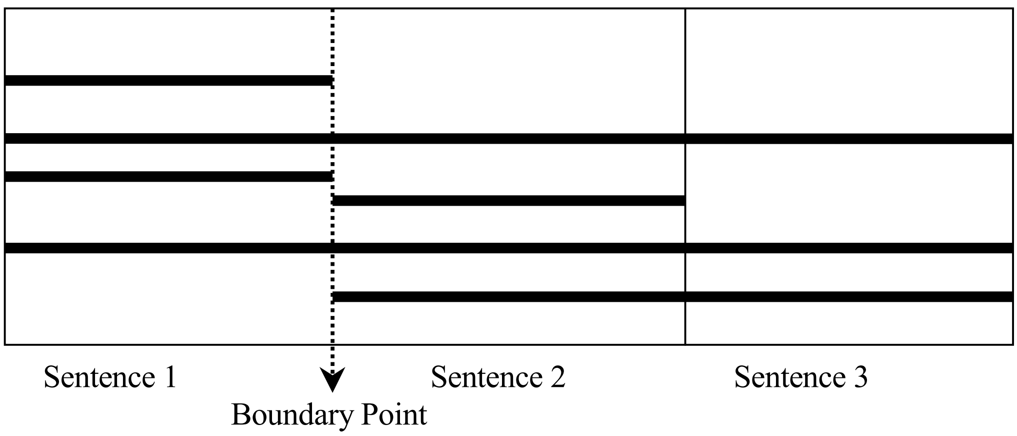


Figure 1. Chain span schema with boundary point detected at end of sentence 1. *w(n, n+1)* values for each of these points are *w(1, 2)* = (2*2) = 4 and *w(2, 3)* = (1*0) = 0.

When all boundary strengths between adjacent sentences 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.

## 6. Experimental Methodology

In this section we present an evaluation of our segmenter SeLeCT. We discuss the evaluation metrics and our decisions regarding the choice of segments making up the corpus.

[2] Variations of our boundary score function were experimented with e.g. summation, weighted summation, and weighted product - of chain begin and end point counts. The above boundary scoring function was chosen as it yields a high level of boundary recall with an acceptable level of precision (see Table 1).

### *6.1 Corpus*

In most test collections used as input to segmentation algorithms a lot of time and effort is spent gathering human annotations i.e. human judged topic shifts. The problem with these annotations lies in determining their reliability since human judges are notoriously inconsistent in their agreement on the beginning and end points of subtopic boundaries [18].

A different approach to segmentation evaluation is available to us in our experiment due to the nature of the segments that we wish to detect. By concatenating distinct CNN broadcast news stories from various CNN news programmes (Night Time, DayLight, International News etc.) and using this as our test set, we eliminate subjectivity from our boundary judgements. So in this case a boundary can be explicitly defined as the joining point between two news stories, in contrast with other test collections, which contain (disjoint) lengthy articles consisting of many subjective subtopic segments. For example, Hearst [6] originally evaluated her TextTiling algorithm on thirteen 'Stargazer' magazine articles which satisfied a certain length criteria (1800 – 2500 words). In comparison each news stories in our collection contains roughly 500 words. Finally our corpus consists of 1001 transcripts, however we only evaluate on the first 1000 boundary end points since finding the end point of the last document in the input stream is a trivial task.

### *6.2 Evaluation Metrics*

We used the standard *precision* and *recall* metrics in our evaluation of text segmentation. These metrics are more commonly used in IR evaluations to measure the ratio between the number of relevant retrieved documents by the IR system and the actual true number of relevant documents (recall) and the number of relevant retrieved documents as a portion of the total number of relevant and non-relevant documents retrieved by the system (precision). Generally speaking the dynamic between these two measures is such that, if the recall of the system is increased (presumably by changing appropriate parameters within the algorithm) the precision of the system will drop and visa versa. In the case of system segmentation evaluation we define recall and precision as follows:

- **Recall** – the number of correctly detected end of news story boundaries as a proportion of the number of actual end of news story boundaries in the test set.
- **Precision** – the number of correctly detected end of news story boundaries as a proportion of the total number of boundaries returned by the segmentation algorithm.

So for example, a segmentation method that correctly classifies 95% of the boundaries it returns might be returning nearly every sentence as a topic boundary and consequently would not be consider a useful segmenter. However determining acceptable precision and recall measures is a subjective decision that depends on the tolerance of the application that required the segments. In our case we intend to use our segments as browsing aids when presenting users in a digital video browsing system developed by DCU [5] with news programmes from a selection of Irish and British news programmes. In this case users will require a high recall scoring system so as to insure that all the news stories discussed are presented to them rather than a situation where two unrelated stories are concatenated together and consequently by not viewing the entire snippet of video segment they may miss an important news story.

It has been noted in text segmentation literature that there are a number of major drawback to using recall and precision metrics in this evaluation domain, since they fail to take in to consideration near boundary misses i.e. if a suggested system boundary is just one sentence away from the true 'end-of-story point' then the system will be penalised just as heavily as a system that has missed the boundary by 10 sentences, obviously a more fatal mistake.

Consequently as Beeferman et al. [9] point out these metrics are insufficiently sensitive for use when trying to find system parameters that yield optimal performance. They try to address these problems by proposing a probabilistic metric that aims to incorporated gradations of segmentation accuracy in terms of insertion (falsely detected segments), deletions (missed segments) and substitutions (close but not exact boundaries)[3]. This idea of considering 'fuzzy boundaries' or substitutions in the evaluation metrics was first proposed by Passoneau and Litman [18], who gave concession to certain boundaries that showed high levels of human annotator disagreement. We base our evaluation methodology on work by Reynar [8] who considers a system boundary correct if it exists within a certain fixed-window of allowable error. Precision and recall measures in the next section are based on observations resulting from increases made to a window of error tolerance.

## 7. Results

Table 1 shows optimal results obtained for the three segmentation systems that took part in our evaluation:

1. **The benchmark system**: that randomly returns boundary positions i.e. results represent a lower bound on performance.
2. **The SeLeCT system**: A lexical chaining approach to segmentation described in detail in Section 5.
3. **TextTiling**: The version of TextTiling we use in this experiment is JTextTile [19], a more efficient java implementation of Hearst's algorithm. Although Hearst recommends window size = 120 and pseudo sentence size = 20, optimal results were achieved using system parameters: window size = 500 and pseudo-sentence size = 20.

Table 1: Precision and Recall values from segmentation on concatenated CNN news stories.

| | ***SeLeCT*** | | ***JtextTile*** | | ***Random Segmentation*** | |
|---|---|---|---|---|---|---|
| ERROR | Recall | Precision | Recall | Precision | Recall | Precision |
| +/-0 | 62.7 | 36.6 | 19.7 | 13.3 | 7.1 | 7.1 |
| +/-1 | 69.2 | 40.4 | 61.6 | 41.5 | 18.4 | 18.4 |
| +/-2 | 77.4 | 45.2 | 79.9 | 53.8 | 29.4 | 29.4 |
| +/-3 | 81.3 | 47.5 | 88.4 | 59.5 | 39.1 | 39.1 |
| +/-4 | 84.3 | 49.2 | 94.1 | 63.4 | 45.9 | 45.9 |
| +/-5 | 85.4 | 49.9 | 96.2 | 64.8 | 51.5 | 51.5 |
| +/-6 | 87.4 | 51.0 | 97.3 | 65.5 | 55.7 | 55.7 |
| +/-7 | 88.3 | 52.7 | 97.8 | 65.9 | 59 | 59 |
| +/-8 | 89.3 | 52.1 | 98.2 | 66.1 | 62.4 | 62.4 |
| +/-9 | 89.9 | 52.5 | 98.3 | 66.2 | 64 | 64 |

[3] This evaluation metric was used in the official TDT 1 segmentation evaluation, which calculates segmentation accuracy in two ways. Firstly a direct evaluation of segmentation is calculated in terms the systems ability to detect story boundaries. Secondly segmentation is evaluated indirectly by measuring event tracking performance on stories return by the segmentation process [4].

Results from Table 1 show variations in precision and recall values using the following error function where *s* is a system boundary point, *b* is an actually boundary point and *n* is the distance in sentences between the actual boundary *b* and the system boundary *s*.

$f(x) = 1 \qquad \text{if } s \ +/- [0 - n] = b$
$f(x) = 0 \qquad \text{Otherwise}$

For example in the case of $n = 5$ if the system boundary is $s = 7$ then the 'correctly detect boundary' score will be incremented if an actually boundary *b* exists between sentence numbers in the range from 2 to 12. The only stipulation on this increment is that sentences boundaries may only be detect once, which takes care of the case where a system boundary might match more than one boundary when the value of *n* is high.

From Table 1 we also observe that our lexical chaining based segmenter SeLeCT significantly outperforms both our benchmark and JTextTile systems. Note that precision and recall values are equivalent in the case of the random segmenter since the segmenter was asked to produce a 1000 boundary points i.e. the actual number of possible topic boundaries in the test set. With regard to the JTextTile performance we see that it shows a poor ability to detect exact boundary points, however as the value of *n* [error tolerance] increases a significant improvement in results is visible. In contrast the SeLeCT system shows high precision and recall values for exact match boundaries and a much slower rate of increase in system performance as boundaries in the immediate vicinity of actual boundaries are consider. This dramatic increase in JTextTile system performance give us an indication as to the style of segmentation returned by the system, where JTextTile has trouble returning the exact location of the end of a news story but it seems to be a excellent indicator of the general area of the actual topic shift. In contrast SeLeCT's segmentation style is an 'all or nothing approach' where the system can correctly pin point a good deal of boundaries but makes few boundary suggestions outside this point. In conclusion then, system performance for both systems is relatively similar when tolerance levels are increased from 0 to 1 for the JTextTile system. However the SeLeCT system is overall more effective as it exhibits the highest performance of the two systems at zero error tolerance. We concur with Hearst's [6] suggestion that significant gains can be achieve if the TextTiling approach was complemented with some auxiliary domain-specific information like cue phrase information, as in turn would the SeLeCT system.

## 8. Conclusions

In this paper we have presented a lexical cohesion based approach to course-grained segmentation of CNN news transcripts resulting in the detection of distinct stories. We have shown that the performance of the SeLeCT system exceeds that of the JTextTile system when exact match story boundaries are required. The next step in our research is to re-evaluate this technique in a real news stream environment. We expect similar high levels of segmentation accuracy will be more difficult to replicate, as closed caption transcripts (in our case teletext) are less informative than CNN speech transcripts. News subtitles are effectively summaries of the audio content of news programme and are dependent on visual cues like speaker change to be fully understood. This lose in information will reduce the internal cohesive strength within stories making subtopics within these stories appear less related than they actually are i.e. this favours fine-grained segmentation. One possible solution to this problem would be to uses a technique like LCA (see Section 4) to expand the context surrounding the teletext based on past news stories.

However we are still optimistic about the success of our current approach, as lexical chains when used for segmentation do not rely on the presence of any explicit cue information, textual structure or training data. They are also domain independent and provide a hierarchical segmentation structure (allowing links between similar segments to be made) if necessary, though linear segmentation was only discuss in this paper.

## 9. Acknowledgements

The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

## References

[1] S. J. Green, *Automatically Generating Hypertext By Comparing Semantic Similarity*. University of Toronto, Technical Report number 366, October 1997.

[2] R. Barzilay, M. Elhadad, *Using Lexical Chains for Text Summarization*. In the proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, 1997.

[3] G. Miller et al., *Five papers on WordNet.* Technical report, Cognitive Science Laboratory, Princeton University, 1990.

[4] J. Allan et al., *Topic Detection and Tracking Pilot Study Final Report*. In the proceedings of the DARPA Broadcasting News Transcript and Understanding Workshop, pp. 194-218, 1998.

[5] A. Smeaton, *Content-based access to digital video: the Físchlár system and the TREC Video track*. In the proceedings of MMCBIR 2001 - Multimedia Content-based Indexing and Retrieval, 2001.

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

[7] C. Manning, *Rethinking text segmentation models: An information extraction case study.* Technical report SULTRY-98-07-01, University of Sydney.

[8] J. Reynar, *Topic Segmentation: Algorithms and Applications, Ph.D. thesis, Computer and Information Science*, University of Pennsylvania, 1998.

[9] D. Beeferman, A. Berger, J. Lafferty, *Text Segmentation using exponential models*. In the proceedings of 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2), 1997.

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

[11] J. Ponte, B. Croft, *Text segmentation by topic.* In the proceedings of the first European Conference on research and advanced technology for digital libraries, 1997.

[12] S. Kaufman, *Second Order Cohesion.* In *Proceeding of PACLING-99,* pp. 209-222, University of Waterloo, Ontario, 1999.

[13] J. Reynar, *An automatic method of finding topic boundaries.* In the proceedings of ACL-94.

[14] H. Kozima, *Text segmentation based on similarity between words.* In the proceedings of ACL-93, pp. 286-288, 1993.

[15] H. Schutze, *Ambiguity Resolution in language learning*, CSLI, 1997.

[16] M. Okumura, T. Honda, *Word sense disambiguation and text segmentation based on lexical cohesion*. In Proceedings of COLING-94, volume 2, pp. 755-761, 1994.

[17] M. A. Stairmand, William J. Black, *Conceptual and Contextual Indexing using WordNet-derived Lexical Chains*. In the proceedings of BCS IRSG Colloquium 1997, pp. 47-65, 1997.

[18] R. Passoneau, D. Litman, *Intention based segmentation: Human reliability and correlation with linguistic cues.* In the proceedings of ACL-93, pp. 148-155, 1993.

[19] F. Choi, *JtextTile: A free platform independent text segmentation algorithm*, 1999. http://www.cs.man.ac.uk/~choif