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.

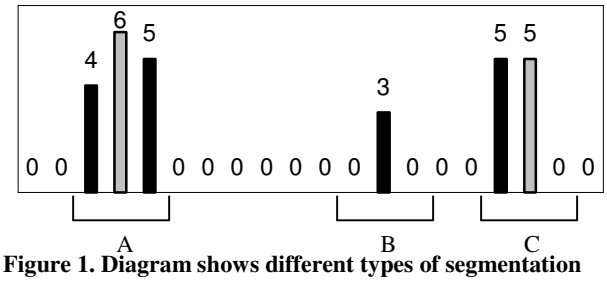

Figure 1. Diagram shows different types of segmentation error; numbers greater than zero are possible boundary positions, while zero scores represent no story boundary point between these two textual units.

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 b