An Analysis of Quantitative Aspects in the Evaluation of Thematic Segmentation Algorithms

Maria Georgescul
ISSCO/TIM, ETI
University of Geneva
1211 Geneva, Switzerland
maria.georgescul@eti.unige.ch

Alexander Clark
Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK
alexc@cs.rhul.ac.uk

Susan Armstrong
ISSCO/TIM, ETI
University of Geneva
1211 Geneva, Switzerland
susan.armstrong@issco.unige.ch

Abstract

We consider here the task of linear thematic segmentation of text documents, by using features based on word distributions in the text. For this task, a typical and often implicit assumption in previous studies is that a document has just one topic and therefore many algorithms have been tested and have shown encouraging results on artificial data sets, generated by putting together parts of different documents. We show that evaluation on synthetic data is potentially misleading and fails to give an accurate evaluation of the performance on real data. Moreover, we provide a critical review of existing evaluation metrics in the literature and we propose an improved evaluation metric.

1 Introduction

The goal of thematic segmentation is to identify boundaries of topically coherent segments in text documents. Giving a rigorous definition of the notion of topic is difficult, but the task of discourse/dialogue segmentation into thematic episodes is usually described by invoking an "intuitive notion of topic" (Brown and Yule, 1998). Thematic segmentation also relates to several notions such as speaker's intention, topic flow and cohesion.

Since it is elusive what mental representations humans use in order to distinguish a coherent text, different surface markers (Hirschberg and Nakatani, 1996; Passonneau and Litman, 1997) and external knowledge sources (Kozima and Furugori, 1994) have been exploited for the purpose of automatic thematic segmentation. Halliday and Hasan (1976) claim that the text meaning is realised through certain language resources and they refer to these resources by the term of cohesion. The major classes of such text-forming resources identified in (Halliday and Hasan, 1976) are: substitution, ellipsis, conjunction, reiteration and collocation. In this paper, we examine one form of lexical cohesion, namely lexical reiteration.

Following some of the most prominent discourse theories in literature (Grosz and Sidner, 1986; Marcu, 2000), a hierarchical representation of the thematic episodes can be proposed. The basis for this is the idea that topics can be recursively divided into subtopics. Real texts exhibit a more intricate structure, including 'semantic returns' by which a topic is suspended at one point and resumed later in the discourse. However, we focus here on a reduced segmentation problem, which involves identifying non-overlapping and non-hierarchical segments at a coarse level of granularity.

Thematic segmentation is a valuable initial tool in information retrieval and natural language processing. For instance, in information access systems, smaller and coherent passage retrieval is more convenient to the user than whole-document retrieval and thematic segmentation has been shown to improve the passage-retrieval performance (Hearst and Plaunt, 1993). In cases such as collections of transcripts there are no headers or paragraph markers. Therefore a clear separation of the text into thematic episodes can be used together with highlighted keywords as a kind of 'quick read guide' to help users to quickly navigate through and understand the text. Moreover automatic thematic segmentation has been shown to play an important role in automatic summarization (Mani, 2001), anaphora resolution and dis-

course/dialogue understanding.

In this paper, we concern ourselves with the task of linear thematic segmentation and are interested in finding out whether different segmentation systems can perform well on artificial and real data sets without specific parameter tuning. In addition, we will refer to the implications of the choice of a particular error metric for evaluation results.

This paper is organized as follows. Section 2 and Section 3 describe various systems and, respectively, different input data selected for our evaluation. Section 4 presents several existing evaluation metrics and their weaknesses, as well as a new evaluation metric that we propose. Section 5 presents our experimental set-up and shows comparisons between the performance of different systems. Finally, some conclusions are drawn in Section 6.

2 Comparison of Systems

Combinations of different features (derived for example from linguistic, prosodic information) have been explored in previous studies like (Galley et al., 2003) and (Kauchak and Chen, 2005). In this paper, we selected for comparison three systems based merely on the lexical reiteration feature: TextTiling (Hearst, 1997), C99 (Choi, 2000) and TextSeg (Utiyama and Isahara, 2001). In the following, we briefly review these approaches.

2.1 *TextTiling* Algorithm

The *TextTiling* algorithm was initially developed by Hearst (1997) for segmentation of expository texts into multi-paragraph thematic episodes having a linear, non-overlapping structure (as reflected by the name of the algorithm). TextTiling is widely used as a de-facto standard in the evaluation of alternative segmentation systems, e.g. (Reynar, 1998; Ferret, 2002; Galley et al., 2003). The algorithm can briefly be described by the following steps.

Step 1 includes stop-word removal, lemmatization and division of the text into 'token-sequences' (i.e. text blocks having a fixed number of words).

Step 2 determines a score for each gap between two consecutive token-sequences, by computing the *cosine similarity* (Manning and Schtze, 1999) between the two vectors representing the frequencies of the words in the two blocks.

Step 3 computes a 'depth score' for each token-sequence gap, based on the local minima of the score computed in step 2.

Step 4 consists in smoothing the scores.

Step 5 chooses from any potential boundaries those that have the scores smaller than a certain 'cutoff function', based on the average and standard deviation of score distribution.

2.2 *C99* Algorithm

The *C99* algorithm (Choi, 2000) makes a linear segmentation based on a divisive clustering strategy and the cosine similarity measure between any two minimal units. More exactly, the algorithm consists of the following steps.

Step 1: after the division of the text into minimal units (in our experiments, the minimal unit is an utterance[1]), stop words are removed and a stemmer is applied.

The second step consists of constructing a similarity matrix $S_{m\times m}$, where m is the number of utterances and an element s_{ij} of the matrix corresponds to the cosine similarity between the vectors representing the frequencies of the words in the i-th utterance and the j-th utterance.

Step 3: a 'rank matrix' $R_{m\times m}$ is computed, by determining for each pair of utterances, the number of neighbors in $S_{m\times m}$ with a lower similarity value.

In the final step, the location of thematic boundaries is determined by a divisive top-down clustering procedure. The criterion for division of the current segment B into $b_1, ...b_m$ subsegments is based on the maximisation of a 'density' D, computed for each potential repartition of boundaries as

$$D = \frac{\sum_{k=1}^{m} sum_k}{\sum_{k=1}^{m} area_k},$$

where sum_k and $area_k$ refers to the sum of rank and area of the k-th segment in B, respectively.

2.3 *TextSeg* Algorithm

The *TextSeg* algorithm (Utiyama and Isahara, 2001) implements a probabilistic approach to determine the most likely segmentation, as briefly described below.

The segmentation task is modeled as a problem of finding the minimum cost $\mathcal{C}(\mathcal{S})$ of a segmentation $\mathcal{S}$. The segmentation cost is defined as:

$$\mathcal{C}(\mathcal{S}) \equiv -logPr(\mathcal{W}|\mathcal{S})Pr(\mathcal{S}),$$

[1] Occasionally within this document we employ the term utterance to denote either a sentence or an utterance in its proper sense.

where $\mathcal{W} = w_1 w_2 ... w_n$ represents the text consisting of n words (after applying stop-words removal and stemming) and $\mathcal{S} = S_1 S_2 ... S_m$ is a potential segmentation of $\mathcal{W}$ in m segments. The probability $Pr(\mathcal{W}|\mathcal{S})$ is defined using Laplace law, while the definition of the probability $Pr(\mathcal{S})$ is chosen in a manner inspired by information theory.

A directed graph $\mathcal{G}$ is defined such that a path in $\mathcal{G}$ corresponds to a possible segmentation of $\mathcal{W}$. Therefore, the thematic segmentation proposed by the system is obtained by applying a dynamic programming algorithm for determining the minimum cost path in $\mathcal{G}$.

3 Input Data

When evaluating a thematic segmentation system for an application, human annotators should provide the gold standard. The problem is that the procedure of building such a reference corpus is expensive. That is, the typical setting involves an experiment with several human subjects, who are asked to mark thematic segment boundaries based on specific guidelines and their intuition. The inter-annotator agreement provides the reference segmentation. This expense can be avoided by constructing a synthetic reference corpus by concatenation of segments from different documents. Therefore, the use of artificial data for evaluation is a general trend in many studies, e.g. (Ferret, 2002; Choi, 2000; Utiyama and Isahara, 2001).

In our experiment, we used artificial and real data, i.e. the algorithms have been tested on the following data sets containing English texts.

3.1 Artificially Generated Data

Choi (2000) designed an artificial dataset, built by concatenating short pieces of texts that have been extracted from the Brown corpus. Any test sample from this dataset consists of ten segments. Each segment contains the first n sentences (where $3 \leq n \leq 11$) of a randomly selected document from the Brown corpus. From this dataset, we randomly chose for our evaluation 100 test samples, where the length of a segment varied between 3 and 11 sentences.

3.2 TDT Data

One of the commonly used data sets for topic segmentation emerged from the Topic Detection and Tracking (TDT) project, which includes the task of story segmentation, i.e. the task of segmenting a stream of news data into topically cohesive stories. As part of the TDT initiative several datasets of news stories have been created. In our evaluation, we used a subset of 28 documents randomly selected from the TDT Phase 2 (TDT2) collection, where a document contains an average of 24.67 segments.

3.3 Meeting Transcripts

The third dataset used in our evaluation contains 25 meeting transcripts from the ICSI-MR corpus (Janin et al., 2004). The entire corpus contains high-quality close talking microphone recordings of multi-party dialogues. Transcriptions at word level with utterance-level segmentations are also available. The gold standard for thematic segmentations has been kindly provided by (Galley et al., 2003) and has been chosen by considering the agreement between at least three human annotations. Each meeting is thus divided into contiguous major topic segments and contains an average of 7.32 segments.

Note that thematic segmentation of meeting data is a more challenging task as the thematic transitions are subtler than those in TDT data.

4 Evaluation Metrics

In this section, we will look in detail at the error metrics that have been proposed in previous studies and examine their inadequacies. In addition, we propose a new evaluation metric that we consider more appropriate.

4.1 $P_\mathbf{k}$ Metric

(Passonneau and Litman, 1996; Beeferman et al., 1999) underlined that the standard evaluation metrics of precision and recall are inadequate for thematic segmentation, namely by the fact that these metrics did not account for how far away is a hypothesized boundary (i.e. a boundary found by the automatic procedure) from a reference boundary (i.e. a boundary found in the reference data). On the other hand, it is desirable that an algorithm that places for instance a boundary just one utterance away from the reference boundary to be penalized less than an algorithm that places a boundary two (or more) utterances away from the reference boundary. Hence (Beeferman et al., 1999) proposed a new metric, called P_D, that allows for a slight vagueness in where boundaries lie. More

specifically, (Beeferman et al., 1999) define P_D as follows[2]:
$P_D(ref, hyp) = \sum_{1 \leq i \leq j \leq N} D(i,j)[\delta_{ref}(i,j) \oplus \delta_{hyp}(i,j)]$.
N is the number of words in the reference data. The function $\delta_{ref}(i,j)$ is evaluated to one if the two reference corpus indices specified by its parameters i and j belong in the same segment, and zero otherwise. Similarly, the function $\delta_{hyp}(i,j)$ is evaluated to one, if the two indices are hypothesized by the automatic procedure to belong in the same segment, and zero otherwise. The $\oplus$ operator is the XNOR function 'both or neither'. $D(i,j)$ is a "distance probability distribution over the set of possible distances between sentences chosen randomly from the corpus". In practice, a distribution D having "all its probability mass at a fixed distance k" (Beeferman et al., 1999) was adopted and the metric P_D was thus renamed P_k.

In the framework of the TDT initiative, (Allan et al., 1998) give the following formal definition of P_k and its components:

$$P_k = P_{Miss} \cdot P_{seg} + P_{FalseAlarm} \cdot (1 - P_{seg}),$$

where:

$$P_{Miss} = \frac{\sum_{i=1}^{N-k} [\delta_{hyp}(i,i+k)] \cdot [1 - \delta_{ref}(i,i+k)]}{\sum_{i=1}^{N-k} [1 - \delta_{ref}(i,i+k)]},$$

$$P_{FalseAlarm} = \frac{\sum_{i=1}^{N-k} [1 - \delta_{hyp}(i,i+k)] \cdot [\delta_{ref}(i,i+k)]}{\sum_{i=1}^{N-k} \delta_{ref}(i,i+k)},$$

and P_{seg} is the *a priori* probability that in the reference data a boundary occurs within an interval of *k* words. Therefore P_k is calculated by moving a window of a certain width k, where k is usually set to half of the average number of words per segment in the gold standard.

Pevzner and Hearst (2002) highlighted several problems of the P_k metric. We illustrate below what we consider the main problems of the P_k metric, based on two examples.

Let $r(i,k)$ be the number of boundaries between positions i and $i+k$ in the gold standard segmentation and $h(i,k)$ be the number of boundaries between positions i and $i+k$ in the automatic hypothesized segmentation.

- Example 1: If $r(i,k) = 2$ and $h(i,k) = 1$ then obviously a missing boundary should be counted in P_k, i.e. P_{Miss} should be increased.
- Example 2: If $r(i,k) = 1$ and $h(i,k) = 2$ then obviously $P_{FalseAlarm}$ should be increased.

However, considering the first example, we will obtain $\delta_{ref}(i, i+k) = 0$, $\delta_{hyp}(i, i+k) = 0$ and consequently P_{Miss} is not increased. By taking the case from the second example we obtain $\delta_{ref}(i, i+k) = 0$ and $\delta_{hyp}(i, i+k) = 0$, involving no increase of $P_{FalseAlarm}$.

In (TDT, 1998), a slightly different definition is given for the P_k metric: the definition of *miss* and *false alarm* probabilities is replaced with:

$$P'_{Miss} = \frac{\sum_{i=1}^{N-k} [1 - \Omega_{hyp}(i,i+k)] \cdot [1 - \delta_{ref}(i,i+k)]}{\sum_{i=1}^{N-k} [1 - \delta_{ref}(i,i+k)]},$$

$$P'_{FalseAlarm} = \frac{\sum_{i=1}^{N-k} [1 - \Omega_{hyp}(i,i+k)] \cdot [\delta_{ref}(i,i+k)]}{\sum_{i=1}^{N-k} \delta_{ref}(i,i+k)},$$

where:

$$\Omega_{hyp}(i, i+k) = \begin{cases} 1, & if\ r(i,k) = h(i,k), \\ 0, & otherwise. \end{cases}$$

We will refer to this new definition of P_k by P'_k. Therefore, by taking the definition of P'_k and the first example above, we obtain $\delta_{ref}(i, i+k) = 0$ and $\Omega_{hyp}(i, i+k) = 0$ and thus P'_{Miss} is correctly increased. However for the case of example 2 we will obtain $\delta_{ref}(i, i+k) = 0$ and $\Omega_{hyp}(i, i+k) = 0$, involving no increase of $P'_{FalseAlarm}$ and erroneous increase of P'_{Miss}.

4.2 *WindowDiff* metric

Pevzner and Hearst (2002) propose the alternative metric called *WindowDiff*. By keeping our notations concerning $r(i,k)$ and $h(i,k)$ introduced in the subsection 4.1, *WindowDiff* is defined as:

$$\textit{WindowDiff} = \frac{\sum_{i=1}^{N-k} [|r(i,k) - h(i,k)| > 0]}{N-k}.$$

Similar to both P_k and P'_k, *WindowDiff* is also computed by moving a window of fixed size across the test set and penalizing the algorithm misses or erroneous algorithm boundary detections. However, unlike P_k and P'_k, *WindowDiff* takes into account how many boundaries fall within the window and is penalizing in "how many discrepancies occur between the reference and the system results" rather than "determining how often two units of text are incorrectly labeled

[2] Let *ref* be a correct segmentation and *hyp* be a segmentation proposed by a text segmentation system. We will keep this notations in equations introduced below.

as being in different segments" (Pevzner and Hearst, 2002).

Our critique concerning *WindowDiff* is that misses are less penalised than false alarms and we argue this as follows. *WindowDiff* can be rewritten as:
$WindowDiff = WD_{Miss} + WD_{FalseAlarm}$,
where:
$WD_{Miss} = \frac{\sum_{i=1}^{N-k}[r(i,k)>h(i,k)]}{N-k}$,

$WD_{FalseAlarm} = \frac{\sum_{i=1}^{N-k}[r(i,k)<h(i,k)]}{N-k}$.

Hence both misses and false alarms are weighted by $\frac{1}{N-k}$.

Note that, on the one hand, there are indeed (N-k) equiprobable possibilities to have a false alarm in an interval of k units. On the other hand, however, the total number of equiprobable possibilities to have a miss in an interval of k units is smaller than (N-k) since it depends on the number of reference boundaries (i.e. we can have a miss in the interval of k units only if in that interval the reference corpus contains at least one boundary). Therefore misses, being weighted by $\frac{1}{N-k}$, are less penalised than false alarms.

Let B_{ref} be the number of thematic boundaries in the reference data. Let's say that the reference data contains about 20% boundaries and 80% non-boundaries from the total number of potential boundaries. Therefore, since there are relatively few boundaries compared with non-boundaries, a strategy introducing no false alarms, but introducing a maximum number of misses (i.e. $k \cdot B_{ref}$ misses) can be judged as being around 80% correct by the *WindowDiff* measure. On the other hand, a segmentation with no misses, but with a maximum number of false alarms (i.e. $(N-k)$ false alarms) is judged as being 100% erroneous by the *WindowDiff* measure. That is, misses and false alarms are not equally penalised.

Another issue regarding *WindowDiff* is that it is not clear "how does one interpret the values produced by the metric" (Pevzner and Hearst, 2002).

4.3 Proposal for a New Metric

In order to address the inadequacies of P_k and *WindowDiff*, we propose a new evaluation metric, defined as follows:
$Pr_{error} = C_{miss} \cdot Pr_{miss} + C_{fa} \cdot Pr_{fa}$,
where:
C_{miss} ($0 \leq C_{miss} \leq 1$) is the cost of a miss, C_{fa} ($0 \leq C_{fa} \leq 1$) is the cost of a false alarm,
$Pr_{miss} = \frac{\sum_{i=1}^{N-k}[\Theta_{ref\_hyp}(i,k)]}{\sum_{i=1}^{N-k}[\Delta_{ref}(i,k)]}$,

$Pr_{fa} = \frac{\sum_{i=1}^{N-k}[\Psi_{ref\_hyp}(i,k)]}{N-k}$,

$$\Theta_{ref\_hyp}(i,k) = \begin{cases} 1, & if\ r(i,k) > h(i,k) \\ 0, & otherwise \end{cases}$$

$$\Psi_{ref\_hyp}(i,k) = \begin{cases} 1, & if\ r(i,k) < h(i,k) \\ 0, & otherwise. \end{cases}$$

$$\Delta_{ref}(i,k) = \begin{cases} 1, & if\ r(i,k) > 0 \\ 0, & otherwise. \end{cases}$$

Pr_{miss} could be interpreted as the probability that the hypothesized segmentation contains less boundaries than the reference segmentation in an interval of k units[3], conditioned by the fact that the reference segmentation contains at least one boundary in that interval. Analogously Pr_{fa} is the probability that the hypothesized segmentation contains more boundaries than the reference segmentation in an interval of k units.

For certain applications where misses are more important than false alarms or vice versa, the Pr_{error} can be adjusted to tackle this trade-off via the C_{fa} and C_{miss} parameters. In order to have $Pr_{error} \in [0,1]$, we suggest that C_{fa} and C_{miss} be chosen such that $C_{fa} + C_{miss} = 1$. By choosing C_{fa}=C_{miss}=$\frac{1}{2}$, the penalization of misses and false alarms is thus balanced. In consequence, a strategy that places no boundaries at all is penalized as much as a strategy proposing boundaries everywhere (i.e. after every unit). In other words, both such degenerate algorithms will have an error rate Pr_{error} of about 50%. The worst algorithm, penalised as having an error rate Pr_{error} of 100% when $k = 2$, is the algorithm that places boundaries everywhere except the places where reference boundaries exist.

5 Results

5.1 Test Procedure

For the three datasets we first performed two common preprocessing steps: common words are eliminated using the same stop-list and remaining words are stemmed by using Porter's algorithm (1980). Next, we ran the three segmenters described in Section 2, by employing the default values for any system parameters and by letting the

[3] A unit can be either a word or a sentence / an utterance.

systems estimate the number of thematic boundaries.

We also considered the fact that C99 and TextSeg algorithms can take into account a fixed number of thematic boundaries. Even if the number of segments per document can vary in TDT and meeting reference data, we consider that in a real application it is impossible to provide to the systems the exact number of boundaries for each document to be segmented. Therefore, we ran C99 and TextSeg algorithms (for a second time), by providing them only the average number of segments per document in the reference data, which gives an estimation of the expected level of segmentation granularity.

Four additional naive segmentations were also used for evaluation, namely: *no boundaries*, where the whole text is a single segment; *all boundaries*, i.e. a thematic boundary is placed after each utterance; *random known*, i.e. the same number of boundaries as in gold standard, distributed randomly throughout text; and *random unknown*: the number of boundaries is randomly selected and boundaries are randomly distributed throughout text. Each of the segmentations was evaluated with P_k, P'_k and *WindowDiff*, as described in Section 4.

5.2 Comparative Performance of Segmentation Systems

The results of applying each segmentation algorithm to the three distinct datasets are summarized in Figures 1, 2 and 3. Percent error values are given in the figures and we used the following abbreviations: *WD* to denote *WindowDiff* error metric; *TextSeg_KA* to denote the TextSeg algorithm (Utiyama and Isahara, 2001) when the average number of boundaries in the reference data was provided to the algorithm; *C99_KA* to denote the C99 algorithm (Choi, 2000) when the average number of boundaries in the reference data was provided to the algorithm; *N0* to denote the algorithm proposing a segmentation with no boundaries; *All* to denote the algorithm proposing the degenerate segmentation *all boundaries*; *RK* to denote the algorithm that generates a *random known* segmentation; and *RU* to denote the algorithm that generates a *random unknown* segmentation.

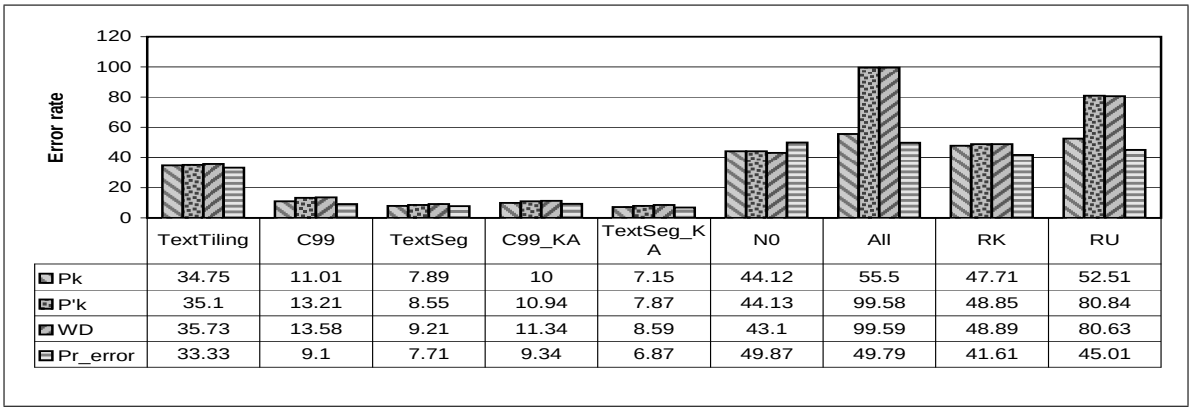
5.2.1 Comparison of System Performance from Artificial to Realistic Data

From the artificial data to the more realistic data, we expect to have more noise and thus the algorithms to constantly degrade, but as our experiments show a reversal of the assessment can appear. More exactly: as can be seen from Figure 1, both C99 and TextSeg algorithms significantly outperformed TextTiling algorithm on the artificially created dataset, when the number of segments was determined by the systems. A comparison between the error rates given in Figure 1 and Figure 2 show that C99 and TextSeg have a similar trend, by significantly decreasing their performance on TDT data, but still giving better results than TextTiling on TDT data. When comparing the systems by Pr_{error}, C99 has similar performance with TextTiling on meeting data (see Figure 3). Moreover, when assessment is done by using *WindowDiff*, P_k or P'_k, both C99 and TextSeg came out worse than TextTiling on meeting data. This demonstrates that rankings obtained when evaluating on artificial data are different from those obtained when evaluating on realistic data. An alternative interpretation can be given by taking into account that the degenerative *no boundaries* segmentation has an error rate of only 30% by the *WindowDiff*, P_k and P'_k metrics on meeting data. That is, we could interpret that all three systems give completely wrong segmentations on meeting data (due to the fact that topic shifts are subtler and not as abrupt as in TDT and artificial data). Nevertheless, we tend to adopt the first interpretation, given the weaknesses of P_k, P'_k and *WindowDiff* (where misses are less penalised than false alarms), as discussed in Section 4.

5.2.2 The Influence of the Error Metric on Assessment

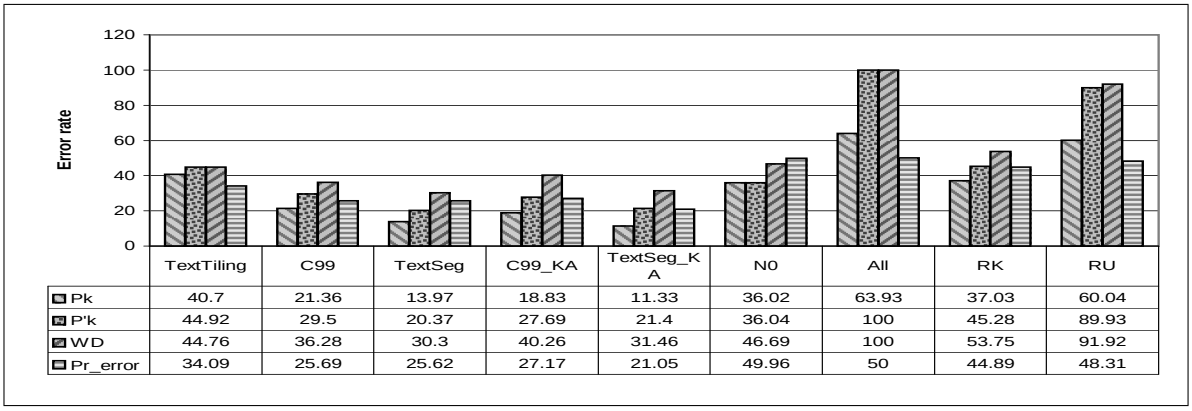
By following the quantitative assessment given by the *WindowDiff* metric, we observe that the algorithm labeled *N0* is three times better than the algorithm *All* on meeting data (see Figure 3), while the same algorithm *N0* is considered only two times better than *All* on the artificial data (see Figure 1). This verifies the limitation of the *WindowDiff* metric discussed in Section 4.

The four error metrics described in detail in Section 4 have shown that the effect of knowing the average number of boundaries on C99 is positive when testing on meeting data. However if we want to take into account all the four error met-



	TextTiling	C99	TextSeg	C99_KA	TextSeg_KA	N0	All	RK	RU
Pk	34.75	11.01	7.89	10	7.15	44.12	55.5	47.71	52.51
P'k	35.1	13.21	8.55	10.94	7.87	44.13	99.58	48.85	80.84
WD	35.73	13.58	9.21	11.34	8.59	43.1	99.59	48.89	80.63
Pr_error	33.33	9.1	7.71	9.34	6.87	49.87	49.79	41.61	45.01

Figure 1: Error rates of the segmentation systems on artificial data, where $k = 42$ and $P_{seg} = 0.44$.



	TextTiling	C99	TextSeg	C99_KA	TextSeg_KA	N0	All	RK	RU
Pk	40.7	21.36	13.97	18.83	11.33	36.02	63.93	37.03	60.04
P'k	44.92	29.5	20.37	27.69	21.4	36.04	100	45.28	89.93
WD	44.76	36.28	30.3	40.26	31.46	46.69	100	53.75	91.92
Pr_error	34.09	25.69	25.62	27.17	21.05	49.96	50	44.89	48.31

Figure 2: Error rates of the segmentation systems on TDT data, where $k = 55$ and $P_{seg} = 0.3606$.

rics, it is difficult to draw definite conclusions regarding the influence of knowing the average number of boundaries on TextSeg and C99 algorithms. For example, when tested on TDT data, C99_KA seems to work better than C99 by P_k and P'_k metrics, while the *WindowDiff* metric gives a contradictory assessment.

6 Conclusions

By comparing the performance of three systems for thematic segmentation on different kinds of data, we address two important issues in a quantitative evaluation. Strong emphasis was put on the kind of data used for evaluation and we have demonstrated experimentally that evaluation on synthetic data is potentially misleading. The second major issue addressed in this paper concerns the choice of a valuable error metric and its side effects on the evaluation assessment.

Acknowledgments This work is supported by the Interactive Multimodal Information Management project (http://www.im2.ch/). Many thanks to Andrei Popescu-Belis and the anonymous reviewers for their valuable comments. We are grateful to the International Computer Science Institute (ICSI), University of California for sharing the data with us. We also wish to thank Michael Galley who kindly provided us the thematic annotations of ICSI data.

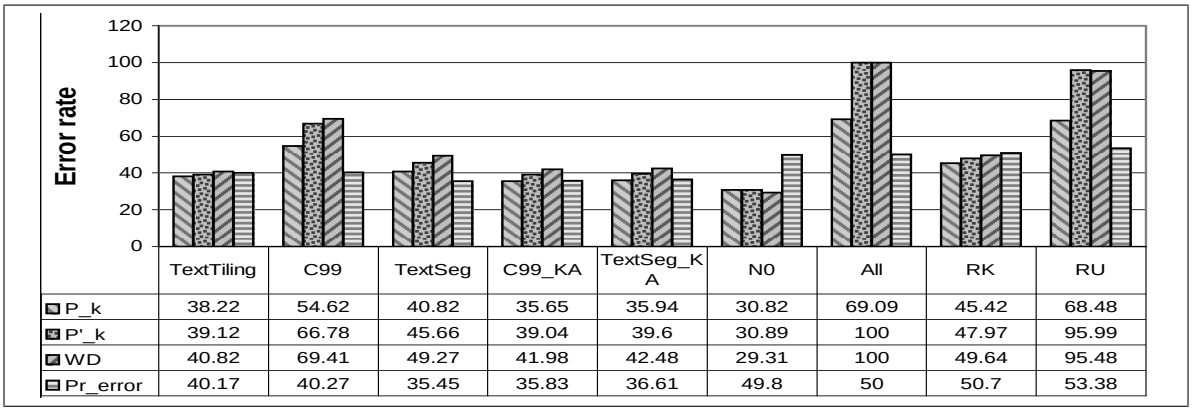
References

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic Detection and Tracking Pilot Study: Final Report. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Landsdowne, VA. Morgan Kaufmann.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34(Special Issue on Natural Language Learning):177–210.

Gillian Brown and George Yule. 1998. *Discourse Analysis*. (Cambridge Textbooks in Linguistics), Cambridge.

Freddy Choi. 2000. Advances in Domain Independent Linear Text Segmentation. In *Proceedings of the 1st*



	TextTiling	C99	TextSeg	C99_KA	TextSeg_KA	N0	All	RK	RU
P_k	38.22	54.62	40.82	35.65	35.94	30.82	69.09	45.42	68.48
P'_k	39.12	66.78	45.66	39.04	39.6	30.89	100	47.97	95.99
WD	40.82	69.41	49.27	41.98	42.48	29.31	100	49.64	95.48
Pr_error	40.17	40.27	35.45	35.83	36.61	49.8	50	50.7	53.38

Figure 3: Error rates of the segmentation systems on meeting data, where $k = 85$ and $P_{seg} = 0.3090$.

Conference of the North American Chapter of the Association for Computational Linguistics, Seattle, USA.

Olivier Ferret. 2002. Using Collocations for Topic Segmentation and Link Detection. In *The 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Michael Galley, Kathleen McKeown, Eric Fosler-Luissier, and Hongyan Jing. 2003. Discourse Segmentation of Multy-Party Conversation. In *Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12:175–204.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Marti Hearst and Christian Plaunt. 1993. Subtopic Structuring for Full-Length Document Access. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, pages 59–68, Pittsburgh, Pennsylvania, United States.

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

Julia Hirschberg and Christine Nakatani. 1996. A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, pages 286 – 293, Santa Cruz, California.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Macias-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. The ICSI Meeting Project: Resources and Research. In *ICASSP 2004 Meeting Recognition Workshop (NIST RT-04 Spring Recognition Evaluation)*, Montreal.

David Kauchak and Francine Chen. 2005. Feature-based segmentation of narrative documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 32–39, Ann Arbor; MI; USA.

Hideki Kozima and Teiji Furugori. 1994. Segmenting Narrative Text into Coherent Scenes. *Literary and Linguistic Computing*, 9:13–19.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Pub Co.

Chris Manning and Hinrich Schtze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, MA, USA.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press Cambridge, MA, USA.

Rebecca J. Passonneau and Diane J. Litman. 1996. Empirical Analysis of Three Dimensions of Spoken Discourse: Segmentation, Coherence and Linguistic Devices.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, 23(1).

Lev Pevzner and Marti Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 16(1):19–36.

Martin Porter. 1980. An Algorithm for Suffix Stripping. *Program*, 14:130 – 137.

Jeffrey Reynar. 1998. *Topic Segmentation: Algorithms and Applications*. Ph.D. thesis, University of Pennsylvania.

TDT. 1998. The Topic Detection and Tracking - Phase 2 Evaluation Plan. Available from World Wide Web: http://www.nist.gov/speech/tests/tdt/tdt98/index.htm.

Masao Utiyama and Hitoshi Isahara. 2001. A Statistical Model for Domain-Independent Text Segmentation. In *ACL/EACL*, pages 491–498.