Topical Segmentation: a Study of Human Performance and a New Measure of Quality

Anna Kazantseva
School of Electrical Engineering
and Computer Science,
University of Ottawa
ankazant@eecs.uottawa.ca

Stan Szpakowicz
School of Electrical Engineering
and Computer Science,
University of Ottawa &
Institute of Computer Science,
Polish Academy of Sciences
szpak@eecs.uottawa.ca

Abstract

In a large-scale study of how people find topical shifts in written text, 27 annotators were asked to mark topically continuous segments in 20 chapters of a novel. We analyze the resulting corpus for inter-annotator agreement and examine disagreement patterns. The results suggest that, while the overall agreement is relatively low, the annotators show high agreement on a subset of topical breaks – places where most prominent topic shifts occur. We recommend taking into account the prominence of topical shifts when evaluating topical segmentation, effectively penalizing more severely the errors on more important breaks. We propose to account for this in a simple modification of the *windowDiff* metric. We discuss the experimental results of evaluating several topical segmenters with and without considering the importance of the individual breaks, and emphasize the more insightful nature of the latter analysis.

1 Introduction

Topical segmentation is a useful intermediate step in many high-level NLP applications such as information retrieval, automatic summarization and question answering. It is often necessary to split a long document into topically continuous segments. Segmentation may be particularly beneficial when working with documents without overt structure: speech transcripts (Malioutov and Barzilay, 2006), newswire (Misra et al., 2011) or novels (Kazantseva and Szpakowicz, 2011). The customary approach is to cast text segmentation as a binary problem: is there a shift of topic between any two adjacent textual units (e.g., sentences or paragraphs)? While necessary, this simplification is quite crude. Topic in discourse usually changes continually; some shifts are subtle, others – more prominent.

The evaluation of text segmentation remains an open research problem. It is a tradition to compile a gold-standard segmentation reference using one or more annotations created by humans. If an automatic segmenter agrees with the reference, it is rewarded, otherwise it is penalized (see Section 4 for details). The nature of the task, however, is such that creating and applying a reference segmentation is far from trivial. The identification of topical shifts requires discretization of a continuous concept – how much the topic changes between two adjacent units. That is why annotators often operate at different levels of granularity. Some people mark only the most prominent topic fluctuations, while others also include finer changes. The task is also necessarily under-defined. In addition to topic changes *per se*, annotators effectively must classify some rhetorical and pragmatic phenomena – exactly how much it is depends on the document genre. For simplicity we do not directly address the latter problem here; we concentrate on the former.

To study how people identify topical shifts in written text, we asked 27 annotators to segment into *episodes* 20 chapters of the novel *The Moonstone* by Wilkie Collins. Each chapter was annotated by 4-6 people. An episode roughly corresponds to a topically continuous segment – the term is defined in Section 3. The analysis of the resulting corpus reveals that while the overall inter-annotator agreement is quite low and is not uniform throughout each chapter. Some topical shifts are marked by most or all annotators, others – by one or by a minority. In fact, only about 50% of all annotated topical shifts are supported by at least 50% of annotators (including near-hits), while the other half is only marked by a minority. In this work we take the agreement about a certain topical shift as a measure of its prominence, and show how this measure can be simply utilized for the purpose of evaluation.

The main claim of this paper is perhaps the following: when evaluating the performance of automatic segmenters, it is important to consider not only the overall similarity between human and machine segmentations, but also to examine the regions of disagreement. When a program misses or misplaces a prominent topic shift – a segment boundary marked by all annotators – it should be penalized more than if it was mistaken about a boundary marked by one person. Similarly, a false positive in the region where none of the annotators found a change in topic is worse than a boundary inserted in a place where at least one person perceived a topic change. We suggest that it is important to use all available reference segmentations instead of compiling them into a single gold standard. We show how a small modification to the popular *windowDiff* (Pevzner and Hearst, 2002) metric can allow considering multiple annotations at once.

To demonstrate the increased interpretive power of such evaluation we run and evaluate several state-of-the art segmenters on the corpus described in this work. We evaluate their performance first in a conventional manner – by combining all available references into one – and then by using the proposed modification. Comparing the results suggests that the information provided by this method differs from what existing methods provide.

Section 2 gives a brief background on text segmentation. Section 3 describes the corpus and how it was collected. Section 4 contain quantitative and qualitative analysis of the corpus and its interpretations. Section 5 proposes a modified version of *windowDiff* and motivates it. Section 6 compares evaluation of three segmenters in several different ways. Section 7 contains the conclusions and outlines directions for future work.

2 Background and Related Work

The goal of topical text segmentation is to identify segments within which the topic under discussion remains relatively constant. A flip-side of this definition is identifying topic shifts – places where the topic shifts significantly or abruptly. In the context of this paper we allow ourselves to use these two definitions interchangeably, sometimes talking about identifying topic shifts, at other times – about identifying topically continuous segments. While the theoretical correctness of such usage remains questionable, it is sufficient for the purpose of our discussion, and it is in line with the literature on the topic.

There is a number of corpora annotated for the presence of topical shifts by one or more annotators. Passonneau and Litman (1997) describe an experiment where seven untrained annotators were asked to find discourse segments in a corpus of transcribed narratives about a movie. While the authors show that the agreement is significant, they also remark that people included segment boundaries at different rates.

Gruenstein, Niekrasz, and Purver (2005) describe the process of annotating parts of two corpora of meeting transcripts: ICSI (Janin et al., 2003) and ISL (Burger, MacLaren, and Yu, 2002). Two people annotated the texts at two levels: major and minor, corresponding to the more and less important topic shifts. Topical shifts were to be annotated so as to allow an outsider to glance at the transcript and get the gist of what she missed. Not unlike our work, the authors report rather low overall inter-annotator agreement. Galley et al. (2003) also compiled a layer of annotation for topical shifts for part of the ICSI corpus, using a somewhat different procedure with three annotators. Malioutov and Barzilay (2006) created a corpus of course lectures segmented by four annotators, noting that the annotators operated at different levels of granularity. In these three projects, manual annotations were compiled into a single gold standard reference for use in evaluating and fine-tuning automatic segmenters.

The work described in this paper is different in several ways. To the best of our knowledge, this is the first attempt to annotate literary texts for topical shifts. Because we collected relatively many annotations for each chapter (four to six), we can make some generalizations as to the nature of the process. In addition to compiling and describing the corpus, we analyze disagreement patterns between annotators. We claim that even though the annotators may not agree on granularity, they do agree at some level, at least with respect to most prominent breaks. We propose that instead of compiling a single reference from multiple annotations it may be more useful to evaluate automatic segmenters against several annotations at once. We will show how to do that.

3 The Overview of the Corpus

Our current work on text segmentation is part of a larger project on automatic summarization of fiction, which is why we chose a XIX century novel, *The Moonstone* by Wilkie Collins, as the text to be annotated. We used two chapters for a pilot study and then another 20 for the large-scale experiment. The annotators worked with individual chapters and were required to align segment boundaries with paragraph breaks.

Objectives. The main question behind this study was this: "How do people identify topical shifts in literature?" This vague question can be mapped to several more specific objectives. First, we sought to verify that topical segmentation of literature was a sensible task from the viewpoint of an untrained annotator. Next, it was important to examine inter-annotator agreement to make sure that the annotators in fact worked on the same phenomena and that the resulting corpus is a reasonable approximation of how people segment literature in general. Third, in addition to analyzing the overall agreement we also took a close look at the type of common disagreements, in search of patterns and insights to evaluate automatic segmenters.

Subjects. The participants were undergraduate university students of an English department at the University of Ottawa. They were recruited by email and received 50 dollars each for their participation. Everyone had to annotate four chapters from *The Moonstone*, not necessarily consecutive ones. The chapters were divided so as to ensure an approximately equal workload.

We had planned six independent annotations for each chapter of the novel.[1] The annotators were divided into five groups, each group asked to read and annotate four distinct chapters. In the end we had three groups with six people, one group with five and one group with four.

Procedure. The experiment was conducted remotely. The students received email packages with detailed instructions and an example of a segmented chapter from a different novel. They had two weeks to annotate the first two chapters and then two more weeks to annotate another two chapters.

The annotators were instructed to read each chapter and split it into episodes – topically continuous spans of text demarcated by the most perceptible shifts of topic in the chapter. We asked the annotators to provide a brief one-sentence description of each episode, effectively creating a chapter outline. The students were also asked to record places they found challenging and to note the time it takes to complete the task.

Because even short chapters of most traditional novels are rather lengthy, we chose to use paragraphs as the basic unit of annotation (sentences are more common in text segmentation literature).

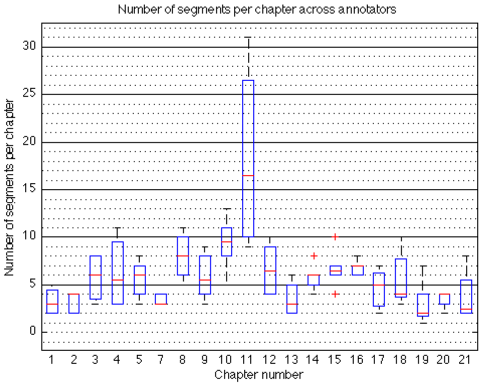
4 Corpus Analysis

Time. On average, an annotator required 137.9 minutes to complete both tasks. The standard deviation was σ = 98.32 minutes appropriately reflecting the fact that some students are very fast readers and besides have already read the novel in one of their classes, while others are quite slow.

The average chapter length is 53.85 paragraphs (σ = 29.31), the average segment length across all annotators is 9.25 paragraphs (σ = 9.77). On average the annotators identified 5.80 episodes (σ = 2.45) per chapter. Figure 1 shows the distribution of the number of segments identified in each chapter. An individual box plot is compiled using all available annotations for the that chapter – six for most, four or five for several. Since the data are plotted for individual chapters, the only source of variance is the disagreement between annotators as to what is the appropriate level of detail for the task. Figure 1 confirms the findings by other researchers that people find topical shifts at different levels of granularity (e.g., (Malioutov and Barzilay, 2006; Gruenstein, Niekrasz, and Purver, 2005)). We take this investigation further and explore whether there are patterns to this disagreement and how they can be interpreted and leveraged.

[1] We hired 30 students. Three did not complete the task.

Figure 1: Distribution of segment counts across chapters.



4.1 Inter-annotator Agreement

In order to make sure that our guidelines are sufficiently clear and the annotators in fact annotate the same phenomenon, it is important to measure inter-annotator agreement (Artstein and Poesio, 2008). This is particularly important given the fact that the resulting corpus is intended as a benchmark dataset for evaluation of automatic segmenters.

When looking at inter-annotator agreement independently of the domain, the most commonly used metrics are coefficients of agreement – α (Krippendorff, 2004), κ (Cohen, 1960; Shrout and Fleiss, 1979), π (Scott, 1955) and several others. In this work we use a multi-annotator version of π, also known in the CL community as Fleiss's κ (Shrout and Fleiss, 1979; Siegel and Castellan, 1988) .

Fleiss's κ is computed as follows:

$$\kappa = \frac{Agreement_{observed} - Agreement_{expected}}{1 - Agreement_{expected}} \tag{1}$$

$$Agreement_{observed} = \frac{1}{ic(c-1)} \sum_{i \in I} \sum_{k \in K} n_{ik}(n_{ik} - 1) \tag{2}$$

$$Agreement_{expected} = \frac{1}{(ic)^2} \sum_{k \in K} n_k^2 \tag{3}$$

where *i* is the number of items to be classified in set *I*, *k* is the number of available categories in set *K*, *c* is the number of annotators, n_{ik} is the number of annotators who assign item *i* to category *k*, n_k is the total number of items assigned to category *k* by all annotators (Artstein and Poesio, 2008, pp. 562-563). Effectively κ measures how much the annotators agree above what can be expected by chance. The value of κ is 0 where there is no agreement above chance and 1 where the annotators agree completely.

While we report κ values for our dataset, it is important to note that κ is ill-suited to measuring agreement in segmentation. The main problem is its insensitivity to near-hits. When asked to segment a document, the annotators often disagree about the exact placement of the boundary but agree that there is a boundary somewhere in the region (e.g., consider paragraphs 9-11 in segmentations in Figure 2). It is desirable to give partial credit to such near-hits instead of dismissing them as utter disagreement. This cannot be achieved with κ. The second problem is the independence assumption: the label for each item must be independent from the labels of all other items. In our case, this would amount to claiming, highly unrealistically, that the probability of a topical shift between two sentences is independent of the topical landscape of the rest of the document.

Two other commonly used agreement metrics are *Pk* (Beeferman, Berger, and Lafferty, 1999) and *windowDiff* (Pevzner and Hearst, 2002), both designed to compare a hypothetical segmentation to a reference, not to measure agreement *per se*. A common feature of both metrics is that they award partial credit to near-hits by sliding a fixed-length window through the sequence and comparing the reference segmentation and hypothetical segmentation at each window position. The window size is generally set at half the average segment length.

Pk (Equation 4) measures the probability that two units randomly drawn from a document are correctly classified as belonging to the same topical segment. *Pk* has been criticized for penalizing false negatives less than false positives and for being altogether insensitive to certain types of error; see (Pevzner and Hearst, 2002, pp. 22-26) for details. Despite its shortcomings, *Pk* is widely used. We report it for comparison with other corpora.

$$Pk(ref, hyp) = \sum_{1 \le i \le j \le n} D(i,j)(\delta_{ref}(i,j) \; XNOR \; \delta_{hyp}(i,j)) \tag{4}$$

Functions δ_{hyp} and δ_{ref} indicate whether the two segment endpoints *i* and *j* belong to the same segment in the hypothetical segmentation and reference segmentation respectively.

windowDiff was designed to remedy some of *Pk*'s shortcomings. It counts erroneous windows in the hypothetical sequence normalized by the total num-

Table 1: Overview of inter-annotator agreement.

	Mean	Std. dev.
κ	0.29	0.15
Pk	0.33	0.17
windowDiff	0.38	0.09

ber of windows. A window is judged erroneous if the boundary counts in the reference segmentation and hypothetical segmentation differ; that is (|ref - hyp| $\neq$ 0) in Equation 5).

$$winDiff = \frac{1}{N-k}\sum_{i=1}^{N-k}(|ref - hyp| \neq 0) \qquad (5)$$

Both *Pk* and *windowDiff* produce penalty scores between 0 and 1, with 1 corresponding to all windows being in error, and 0 – to a perfect segmentation.

Table 1 reports *Pk*, *windowDiff* and κ values for our corpus. *Pk* and *windowDiff* are computed pairwise for all annotators within one group and then averaged. We set the window size to half the average segment length as measured across all annotators who worked on a given chapter. The values are computed for each group separately; Table 1 shows the averages across five groups.

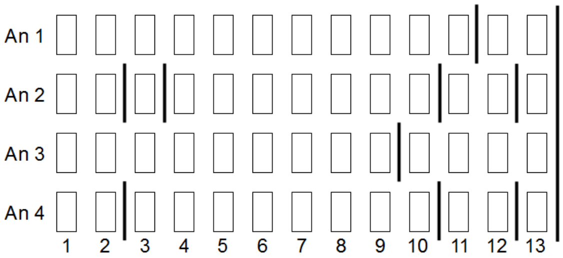
Even by most relaxed standards, e.g., (Landis and Koch, 1977), the κ value of 0.38 corresponds to low agreement. This is not surprising, since it only includes the cases when the annotators agree exactly where the boundary should be. For the purpose of our task, such a definition is too strict.

The values of *windowDiff* and *Pk* are more reasonable; *windowDiff* = 0.34 means that on average a pair of annotators disagrees on 34% of windows. *windowDiff* was originally designed to compare only two segmentations. Our strategy of computing its values pairwise is perhaps not optimal but in the absence of another metric allowing to account for near-hits we are practically forced to use it as a primary means of inter-annotator agreement.

4.2 Patterns of Disagreement

Figure 2 shows the segmentation of the shortest chapter in the dataset. The overall agreement is quite low (*windowDiff*=0.38, $\kappa = 0.28$). This is not surprising, since annotators 1 and 3 found two segments, annotator 3 – five segments, and annotator 4 – four. Yet all annotators agree on certain things: everyone found that there was a significant change of topic between paragraphs 9 and 11 (though they disagree on its exact placement). It is therefore likely that the topical shift between paragraphs 9 and 11 is quite prominent. Annotators 2 and 4 chose to place a segment boundary after paragraph 2, while annotators 1 and 3 did not place one there. It is likely that the topical shift occurring there is less prominent, although perceptible. According to these annotations, the least perceptible topic shifts in the chapter occur after paragraph 4 (marked only by annotator 2) and possibly after paragraph 11 (marked only by annotator 1). Overall, glancing at these segmentations suggests that there is a prominent topical shift between paragraphs 9-11, three significant ones (after 2, 10 and 12) and several minor fluctuations (after 3 and possibly after 10 and 11).

Figure 2: Example segmentation for Chapter 1.



Looking at the segmentations in Figure 2 it seems likely that the disagreements between annotators 2 and 4 are due to granularity, while the annotators 1 and three disagree more fundamentally on where the topic changes. When measuring agreement, we would like to be able to distinguish between disagreements due to granularity and disagreements due to true lack of agreement (annotator 1 and 3). We would also like to leverage this information for the evaluation of automatic segmenters.

Distinguishing between true disagreement and different granularity while taking into account near-hits is not trivial, especially since we are working with multiple annotations simultaneously and there is no *one* correct segmentation.

In order to estimate the quality of individual boundaries and look inside the segmented sequence, we approximate the quality of each suggested segment boundary by the percentage of annotators who marked it. Since the annotators may disagree on the exact placement of the boundaries, our measurement must be relaxed to allow for near-hits.

Figure 3: Quality of segment boundaries.

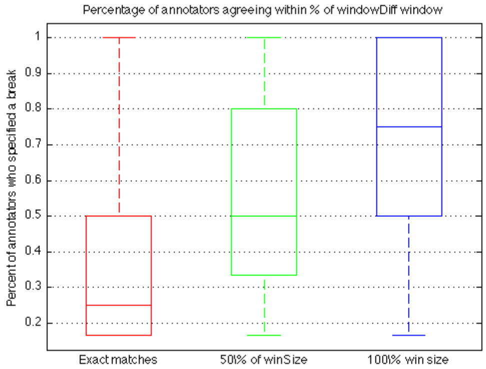


Figure 3 shows the distribution of segment boundaries using three different standards of quality. We consider all segment boundaries introduced by at least one annotator. Then, for each suggested boundary we compute how much support there is from peer annotators: what percentage of annotators included this boundary in their segmentation. The leftmost box plot in Figure 3 corresponds to the most strict standard. When computing support we only consider perfect matches: segment boundaries specified in exactly the same location (window size = 0). The middle box plot is more relaxed: we consider boundaries found within half of a *windowDiff* window size of the boundary under inspection. The rightmost box plot corresponds to the inclusion of boundaries found within a full *windowDiff* window size of the boundary under inspection.

Looking at exact matches (the leftmost box plot), we observe that at least a half of segment boundaries were specified by less than 25% of annotators (which corresponds to one person). It explains why κ values in Table 1 are so low: this is the only sort of agreement κ captures. Also one can notice that at most 25% of the boundaries have the support of more than 50% of the annotators.

The picture changes if we consider all boundaries within a tight window around the candidate boundary (the middle box plot). This standard is twice as strict as the regular *windowDiff* evaluation. Here 50% of all boundaries are marked by at least 35% at and most 80% of annotators. Only 25% of boundaries are marked by less than 30% of the annotators.

The rightmost plot looks even better. If we consider the support found within a window size of any candidate boundary, then 50% of all boundaries are supported by over 70% of annotators. However, we find this way of measuring support too optimistic. The reason is, again, the difference in the granularity of segmentations. The window size used for these measurements is based on the average segment length across all annotations. For example, the average segment length for segmentation shown in Figure 2 is 4, making the window size 2. This size is too relaxed for annotators 2 and 3, who were very detailed. Due to the excessively large window there will almost always be a boundary where fine-grained annotations are concerned, but those boundaries will not correspond to the same phenomena. That is why we think that a stricter standard is generally more appropriate. This is especially the case since we are working with paragraphs, not sentences. A distance of 2-3 sentences is quite tolerable, but a distance of 2-3 paragraphs is considerable, and it is far more likely that a stricter notion of near-hits needs to be considered.

5 Proposed Modification to *windowDiff*

WindowDiff compares two segmentations by taking into account near-hits – penalizing them proportionally to how far a hypothetical segment boundary is from a reference boundary. Section 4.2 argued that some boundaries are more prominent. We aim to modify *windowDiff* so the prominence of the boundaries matters in evaluating automatic segmenters.

Recall that to compute *windowDiff* we slide a window through the reference and the hypothetical segmentation and check whether the number of boundaries is equal at each window position. The number of erroneous windows is then normalized:

$$winDiff = \frac{1}{N-k}\sum_{i=1}^{N-k}(|ref_i - hyp_i| \neq 0) \qquad (6)$$

ref_i and $hypo_i$ are the counts of boundaries in a given window in the reference and the hypothetical sequence, *N* is the length of the complete sequence, *k* is the window size (so there are *N* - *k* windows).

The prominence of a boundary can be approximated by how many annotators specified it in their segmentations. One simple way to take prominence into account is to slide a window through all available segmentations, not just one. A straighforward modification to equation (6) achieves that:

$$winDiff' = \frac{1}{h(N-m)} \sum_{a=1}^{h} \sum_{i=1}^{N-m} (|ref_{ai} - hyp_i| \neq 0) \quad (7)$$

A is the set of all available annotations and h is their total number. Effectively, for each position of the window the hypothetical output is penalized as many times as there are reference annotations with which it disagrees. Note that the window size *m* is different from that used for pair-wise comparisons. Following the convention, we recommend setting it to half of the size of an average segment length (averaged over all available references). The size of the window effectively specifies a tolerance threshold for what is an acceptable near-hit (as opposed to a plain miss), and can be modified accordingly.

windowDiff and *Pk* range from 0 to 1, with 0 corresponding to an ideal segmentation. The upper and lower bounds for Equation 7 are different and depend on how much the reference segmentations agree between themselves.[2]

Let us refer to the most popular opinion for a given position of the window as the *majority opinion*. Then, for each window, the smallest possible penalty is assigned if the hypothetical segmentation correctly "guesses" the majority opinion (the window then receives a penalty equal to the number of annotators disagreeing with the majority opinion):

$$best\_case = \frac{1}{N-m} \sum_{i=1}^{N-m} (h - majority\_support) \quad (8)$$

Here $majority\_support$ is the number of annotators who support the most frequent opinion.

Conversely, to merit the highest penalty, a hypothetical segmentation must "guess" the least popular opinion (possibly an opinion not supported by any annotators) at each window position. In Equation 9, $unpopular\_support$ is the number of annotators who agree with the least popular opinion.

$$worst\_case = \frac{1}{N-m} \sum_{i=1}^{N-m} (h - unpopular\_support) \quad (9)$$

In order to have a multi-annotator version of *windowDiff* interpretable within the familiar $[0, 1]$ interval, we normalize Equation 7:

$$multWinDiff = \frac{(\sum_{a=1}^{h} \sum_{i=1}^{N-m}(|ref_a - hyp| \neq 0)) - best\_case}{h(N-m)(worst\_case - best\_case)} \quad (10)$$

The best and the worst-case bounds serve as indicators of how much agreement there can be between reference segmentations and so as indicators of how difficult to segment a given document is.

The *multWinDiff* metric in Equation 10 has the same desirable properties as the original metric, namely it takes into account near hits and penalizes according to how far the reference and hypothetical boundaries are. Additionally, for each window position it takes into account how much a hypothetical segmentation is similar to all available annotations, thus penalizing mistakes according to the prominence of boundaries (or to the certainty that there are no boundaries). [3]

6 Experiments

In order to illustrate why using a single gold-standard reference segmentation can be problematic, we evaluate three publicly available segmenters, MinCutSeg (Malioutov and Barzilay, 2006), BayesSeg (Eisenstein and Barzilay, 2008) and APS (Kazantseva and Szpakowicz, 2011), using several different gold standards and then using all available annotations. The corpus used for evaluation is *The Moonstone* corpus described in Sections 3-4. We withheld the first four chapters for development and used the remaining 16 for testing. We also compared the segmenters to a random baseline which consisted of randomly selecting a number of boundaries equal to the average number of segments across all available annotations.

None of the segmenters requires training in the conventional sense, but APS and MinCutSeg segmeters come with scripts allowing to fine-tune several parameters. We selected the best parameters for these two segmenters using the first four chapters of the corpus. BayesSeg segmeter, a probabilistic segmenter, does not require setting any parameters.

[2] In our case the upper bound corresponds to the worst-case and the lower bound to the best-case scenario. To avoid confusion, we talk of *the best-case* and *the worst-case bounds*.

[3] Java code to compute *multWinDiff* is available as a part of the APS segmenter. Both the corpus and the software can be downloaded from ⟨`www.site.uottawa.ca/∼ankazant`⟩.

	APS	Bayes	MinCut	Rand.
windowDiff $\geq$50%	0.60.	0.66	0.73	0.73
windowDiff $\geq$30%	0.61	0.52	0.69	0.61
windowDiff union	0.6	0.53	0.63	0.65
windowDiff annotator 1	0.66	0.57	0.74	0.76
windowDiff annotator 4	0.62	0.7	0.69	0.74
windowDiff annotator 2	0.61	0.6	0.66	0.69
multWinDiff	0.23	0.28	0.31	0.41

Table 2: Comparing the three segmenters and a random baseline using different references for computing *windowDiff*. *windowDiff*$\geq$50% - the gold standard consists of all boundaries specified by at least 50% of the annotators; *windowDiff*$\geq$30% – all boundaries specified by at least 30% of the annotators; *windowDiff*union – all boundaries specified by at least one person; *windowDiff*annotator a - comparisons against individual annotators. *multWinDiff* is multi-annotator *windowDiff*from equation (10).

Table 2 sums up the results. Each row corresponds to one reference segmentation and metric – regular *windowDiff* in the first six rows. We compiled several flavours of consensus reference segmentations: 1) all boundaries marked by $\geq$ 50% of the annotators (*windowDiff* $\geq$ 50%), 2) all boundaries marked by $\geq$ 30% of the annotators (*windowDiff* $\geq$ 30%), 3) all boundaries marked by at least one annotator (*windowDiff* union). To illustrate why comparing against a single annotation is unreliable, we report comparisons against three single-person annotations (*windowDiff* annotator 1, 4, 2). *multWinDiff* is the proposed multi-annotator version from Equation 10. The best-case bound for *multWinDiff* is 0.21 and the worst-case bound is 1.0.

Each segmenter produced just one segmentation, so the numbers in the Table 2 differ only depending on the mode of evaluation. The cells are coloured. The lightest shade correspond to the best performance, darker shades – to poorer performance. The actual values for the first six rows are rather low, but what is more bothersome is the lack of consistency in the ranking of segmenters. Only the random baseline remains the worst in most cases. The APS and BayesSeg segmenters tend to appear better than the MinCutSeg but it is not always the case and the rankings among the three are not consistent.

The last row reports multi-annotator *windowDiff* which takes into account all available references and also the best-case and the worst-case bounds. In principle, there is no way to prove that the metric is better than using *windowDiff* and a single reference annotation. It does, however, take into account all available information and provides a different, if not unambiguously more true, picture of the comparative performance of automatic segmenters.

7 Conclusions and Future Work

We have described a new corpus which can be used in research on topical segmentation. The corpus is compiled for fiction, a genre for which no such corpus exists. It contains a reasonable number of annotations per chapter to allow an in-depth analysis of topical segmentation as performed by humans.

Our analysis of the corpus confirms the hypothesis that when asked to find topical segments, people operate at different levels of granularity. We show that only a small percentage of segment boundaries is agreed upon by all or almost all annotators. If, however, near-hits are considered, suggested segment boundaries can be ranked by their prominence using the information about how many people include each boundary in their annotation.

We propose a simple modification to *windowDiff* which allows for taking into account more than one reference segmentation, and thus rewards or penalizes the output of automatic segmenters by considering the severity of their mistakes. The proposed metric is not trouble-free. It is a window-based metric so its value depends on the choice of the window size. While it has become a convention to set the window size to half of the average segment length in the reference segmentation, it is not obvious that the same logic applies in case of multi-annotator *windowDiff*. The metric also hides whether false positives or false negatives are the main source of error.

However, despite all these shortcomings the metric offers an advantage of being able to evaluate hypothetical segmentations with more subtlety than those using a single gold standard reference. When using regular *windowDiff* and a single reference segmentation one is forced to evaluate based on binary comparisons: if a given hypothetical boundary is similar to the gold standard segmentation (e.g., the majority opinion). Divergent segmentations are penalized regardless of whether they are similar to minority opinions (and thus feasible, if less likely) or if they are completely different from anything created by humans (and thus probably genuinely erroneous). However, our version of *windowDiff* takes into account multiple annotations and gives partial reward to segmentations based on how similar there are to any human segmentation, not just the majority opinion (while giving preference to high agreement with the majority opinion).

To evaluate the output of topical segmenters is hard. There is disagreement between the annotators about the appropriate level of granularity and about the exact placement of segment boundaries. The task itself is also a little vague. Just as it is the case in automatic text summarization, generation and other advanced NLP tasks, there is no single correct answer and the goal of a good evaluation metric is to reward plausible hypotheses and to penalize improbable ones. It is quite possible that a better metric than the one proposed here can be devised (e.g., Fournier and Inkpen (2012)). We feel, however, that any reliable metric for evaluating segmentations must – in one manner or another – take into account more than one annotation and the prominence of segment breaks.

References

Artstein, Ron and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Beeferman, Doug, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34:177–210, February.

Burger, Susanne, Victoria MacLaren, and Hua Yu. 2002. The ISL meeting corpus: the impact of meeting type on speech style. In *INTERSPEECH'02*.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales . *Educational and Psychological Measurement*, 20:37–46.

Eisenstein, Jacob and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii, October.

Fournier, Chris and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of NAACL-HLT 2012 (this volume)*, Montréal, Canada, June.

Galley, Michel, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 562–569, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gruenstein, Alexander, John Niekrasz, and Matthew Purver. 2005. Meeting Structure Annotation: Data and Tools. In *In Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 117–127.

Janin, Adam, Don Baron, Jane Edwards, D. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, A.ndreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, volume 1, pages 364–367, April.

Kazantseva, Anna and Stan Szpakowicz. 2011. Linear Text Segmentation Using Affinity Propagation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 284–293, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Krippendorff, Klaus. 2004. *Content Analysis. An Introduction to Its Methodology.* Sage Publications.

Landis, J. Richards and Garry G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Malioutov, Igor and Regina Barzilay. 2006. Minimum Cut Model for Spoken Lecture Segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Sydney, Australia, July.

Misra, Hemant, François Yvon, Olivier Cappé, and Joemon M. Jose. 2011. Text segmentation: A topic modeling perspective. *Information Processing and Management*, 47(4):528–544.

Passonneau, Rebecca J. and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, March.

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

Scott, William. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quaterly*, 19(3):321–325.

Shrout, Patrick E. and Joseph L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Siegel, Sidney and John. N. Jr. Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw Hill, Boston, MA.