

# Pagination: It's What You Say, Not How Long It Takes To Say It

Joshua Hailpern  
HP Labs  
Palo Alto, CA, USA  
joshua.hailpern@hp.com

Niranjan Damera Venkata  
HP Labs  
Chennai, Tamil Nadu, India  
niranjan.damera-venkata@hp.com

Marina Danilevsky  
University of Illinois  
Urbana, IL, USA  
danilev1@illinois.edu

## ABSTRACT

Pagination - the process of determining where to break an article across pages in a multi-article layout - is a common layout challenge for most commercially printed newspapers and magazines. To date, no one has created an algorithm that determines a minimal pagination break point based on the content of the article. Existing approaches for automatic multi-article layout focus exclusively on maximizing content (number of articles) and optimizing aesthetic presentation (e.g., spacing between articles). However, disregarding the semantic information within the article can lead to overly aggressive cutting, thereby eliminating key content and potentially confusing the reader, or setting too generous of a break point, thereby leaving in superfluous content and making automatic layout more difficult. This is one of the remaining challenges on the path from manual layouts to fully automated processes that still ensure article content quality. In this work, we present a new approach to calculating a document minimal break point for the task of pagination. Our approach uses a statistical language model to predict minimal break points based on the semantic content of an article. We then compare 4 novel candidate approaches, and 4 baselines (currently in use by layout algorithms). Results from this experiment show that one of our approaches strongly outperforms the baselines and alternatives. Results from a second study suggest that humans are not able to agree on a single "best" break point. Therefore, this work shows that a semantic-based lower bound break point prediction is necessary for ideal automated document synthesis within a real-world context.

## Categories and Subject Descriptors

I.7.2 [Document Preparation]: Desktop Publishing; I.7.4 [Document and Text Processing]: Electronic Publishing

## Keywords

Pagination; Truncation; Novelty; Semantic; Cut; SLM

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*DocEng'14*, September 16–19, 2014, Fort Collins, Colorado, USA.  
Copyright 2014 ACM 978-1-4503-2949-1/14/09 ...\$15.00.  
<http://dx.doi.org/10.1145/2644866.2644867>.

## 1. INTRODUCTION

Traditional document composition is an iterative process involving copy editors and professional publication designers who work in concert to make decisions on what content to include and how to format it for aesthetic presentation. Within the context of newspapers or magazines, each page can contain multiple articles. However, space constraints dictate that the full text of any (or all) given articles can not be presented on a single page. Articles are therefore broken or paginated across a paper. The first portion of an article is presented upfront, while the full text is presented later on, should the reader wish to consume more. Copy editors and designers work to strike a balance between presenting enough content on this front page so readers can understand the article, while maximizing the space constraints and aesthetic presentation of the page as a whole.

Automated document composition algorithms have been introduced to automate this largely manual workflow. Current approaches to automated document synthesis/layout focus extensively on content and aesthetic maximization [17, 18]. These optimization algorithms work with a series of rules and constraints to maximize the presentation within the spatial and visual constraints.

While these approaches have largely been successful at addressing issues of presentation, they do not take into account the content of the articles. This can result in break points being set too aggressively, resulting in missing key content and a potential misrepresentation of the article content, or too late in the article, potentially lowering the flexibility of layout algorithms by requiring more content to be placed on the front page. To combat these problems, many of these layout algorithms enforce minimal content requirements. However, these rules are arbitrary, consisting of static or relative lower bounds (e.g., at least 2 sentences, or at least 20% of the article). Given that these rules are not related to the *actual* article, key content can still be left out, or excess text may still be forced to be included.

The primary contribution of this work is an algorithm that predicts a lower-bound break point for news article pagination based upon the semantic information contained within the article. Our approach not only outperforms existing solutions, but to the authors' knowledge, is also the first such semantic based pagination algorithm.

In this paper we develop four candidate break point algorithms, all of which are based on the semantic information of an article to be paginated. These four approaches are then directly compared against four baselines currently used in both print and digital media. To facilitate these compar-

isons, we conduct an extensive Mechanical Turk study across 7 subject areas (Sports, US News, Entertainment, etc). Results from this experiment show that one of our approaches strongly outperforms the baselines and alternatives. We also investigate whether an ‘ideal’ break point could be found for paginating news articles. Results from a large CrowdFlower study strongly suggest that humans can not agree on a single best break point. Therefore, this work shows that a semantic-based lower bound break point prediction is necessary for ideal for automated document synthesis within a real-world context.

## 2. RELATED WORK

The majority of work on pagination - the need to split a document or series of documents over multiple pages - focuses on content maximization and aesthetic presentation. Such approaches aim to create ‘high-quality documents,’ defined as documents ‘without unwanted empty areas’ [7]. Within the context of a single document, such as a book or a paper, pagination impacts the layout relationship of text content to related figures or tables [5].

However, the problem complexity greatly increases for automated document layout of newspapers or magazines [12, 17, 18]. In addition to optimizing the placement of figures within text, a given page can also contain content from multiple articles, and all content must fit within a predetermined page count. When determining how to break up articles across multiple pages, the preferred approach is to use constraint-based layout models, in which layout specifications are described by linear constraints imposed on the items within a layout [2, 19, 23, 29, 35]. The focus of these models is on the composition quality [1, 8], and they in no way account for the semantic information contained within the articles themselves.

### 2.1 Automatic Summarization

A similar problem to pagination is that of single document summarization - the process by which the text of an article is reduced either by extraction (lifting sentences from the original text) [24] or abstraction (using natural language processing techniques to generate new sentences) [11]. While these approaches are not used by layout or copy editors, they are relevant to the construction and evaluation of our semantic-based break point techniques.

Corpus summarization uses a large collection of documents to build a model of the topics being discussed (e.g. topic modeling [4, 6], SumBasic [13], KLSum [26]) or opinions rendered (e.g. Opinion Mining [15]). Corpus summarization approaches rely upon a large body of documents (e.g., a collection of tweets [6]) from which patterns about the ‘whole’ can be derived, and are generally easier and more powerful because they have more data from which to draw summaries. In contrast, single document summarization [22, 21] utilizes only one document to create a summary. Within single document summarization, most algorithms are designed to summarize long (e.g. book) [34], well structured (e.g. chapters or sections) text [9, 30, 33], thus maximizing the amount of text and structural cues from which to derive summaries. The most notable exceptions to single unstructured documents are TextRank [24] and LexRank [10]. Both of these algorithms use a simple graph-based approach, treating each sentence as a node. The summary sentence of the document is calculated by finding the centroid of the graph based on a

distance vector. There has been some work specifically with summarizing news articles by extracting the most important facts from the article [20]. Finally, it has shown that using the first  $n$  sentences of a news article as the article’s summary performs very well [25], which adds motivation to our end goal.

### 2.2 Document Sub-Topic Segmentation

Another tangential, though different, problem space is that of document sub-topic segmentation, most notably TextTiling [16, 27]. The main focus of Sub-Topic Segmentation is to divide a document into subtopics, or sections. These sections are not ranked, as to importance or quality, but strive to focus on conceptual shifts in the content being discussed. TextTiling is a straightforward and simplistic [3] method that examines lexical co-occurrence of terms between phrases in a document, and identifying sharp changes as breaks or dividers between subtopics. However, because these approaches lack of ranking or quality of each sub-topic shift, their direct application to the pagination problem are limited and a direct comparison of TextTiling with this problem is difficult. However, this work does suggest techniques or broad approaches we can build upon; detecting breaks based on semantic change variation across a document relative to a mean change, using occurrence of vocabulary to signify information/theme [16].

## 3. SCOPING & KEY TERMS

The goal of this work is to create and test a novel algorithm that can use a news article’s semantics (rather than layout/spacing optimization) to determine a lower bound where to break an article for pagination. This serves two key purposes: first, if an article is broken too early, the reader may miss key information and potentially be misled about the article’s content; second, if too much of an article’s content is forced to be shown, the page layout may not be as flexible because more text would be required to be presented therefore, there may not be space to present as many articles on the front page.

While there are many applications of this technique, this work specifically looks at news articles. Our novel approaches draw on and compare with pagination techniques both in print and digital presentation.

In this context, we define the following terms:

**Topic:** Refers to the use in the English vernacular, such as the topic of a news article (e.g. a specific bomb that goes off in a specific country on a specific day), rather as it is used in the context of Topic Modeling approaches.

**Article:** A piece of written text about a specific topic (e.g. a specific New York Times article about a bomb that went off in a country).

**Break Point:** The location at which you ‘break’ or stop an article, such that the content before the break point is displayed on the current page, and the content after the break point goes on a different page, or after a ‘read more’ link.

**Lower Bound Break Point:** The location within a news article denoting that the reader would have, by this time, gained a general understanding about the topic being discussed, but has not

yet seen all the details or nuances. Breaking before this lower bound is likely to cause the reader to miss a key concept or aspect of the article, and risk being misled about the content. However, if more space is available, a layout algorithm can choose to include more text from a given article.

**Pagination:** the task of determining where to break an article across a multi-article, multi-page layout.

**Corpus:** A collection of articles on the same topic (e.g. 100 news articles about a specific bombing event).

**Subject:** An overarching thematic grouping of articles or corpuses (e.g. Sports, Politics)

**Document:** Any set of sentences, which could be a single article, a subset of an article, multiple articles, a Corpus or Subject, etc.

**Semantics:** Refers to the english definition of “semantics,” focusing on the *meaning* of words or phrases.

**It should be noted that there is an important distinction between pagination (the focus of this paper) and truncation (not discussed in this paper – see [14]).** Pagination, as used in this context, implies that after the break point, a reader will see more of the given article, usually on a subsequent page. On the contrary, when an article is truncated, the portion of the article which follows the break point will never be presented to the reader in any form. We do *not* deal with truncation tasks in this work. Further, this work does not perform document summarization, or sub-topic identification. These approaches can be used to inform our design, but the functionally address and solve a different problem and their output is not directly comparable to the output of pagination.

In this paper, we first describe the algorithms developed for content-aware pagination, as well as the baselines for comparison. Next, we present our experiment comparing the quality of the suggested break points based on article content. Finally, we briefly explore the challenges of predicting an ‘ideal’ break point in a document.

## 4. LOWER BOUND ALGORITHMS

The primary goal of this works is to develop and compare algorithms that can predict a lower bound break point for pagination of news articles. This section details the four content-agnostic algorithms that are used as ‘current state’ baselines, as well as our novel predictive algorithms that directly leverage the semantic content of the articles.

### 4.1 Four Baselines

To facilitate a comparison, we leveraged four baselines that are commonly used as lower-bounds for pagination: the first sentence, the first two sentences, the first paragraph, and the first twenty percent of the article.

#### 4.1.1 One & Two Sentences

Commonly used in online news websites (e.g. Google News, NBCNews.com), the first or first two sentences of a news article are used to ‘preview’ the full text. Thus, pagination occurs after these sentences, and the reader can

consume the full text by clicking a link<sup>1</sup>. We refer to these two techniques as **One Sentence** and **Two Sentences**, respectively.

#### 4.1.2 One Paragraph

While writing, a structuraly delimiter between semantic concepts is often the paragraph break. Thus, the first paragraph of an article can be thought of as the first digestible nugget of an article’s content and may make a natural break point. This is used at Wall Street Journal’s website (wsj.com) for non-subscribers who seek to access subscriber only articles. We will refer to this as **One Paragraph**.

#### 4.1.3 Twenty Percent Rounded Up to Nearest Paragraph

When constructing printable newspapers, many sources use an article length dependent approach. Minimal break-points are calculated by the first 20% of a document (measured by character count), then rounded up to the end of the current paragraph. This approach assumes that within the first fifth of a document, key concepts have been presented to the reader, and are roughly bounded by paragraph delimitation. We refer to this technique as **Twenty Percent**.

## 4.2 Keyword Novelty

The first approach we develop is predicated on the idea that as a reader traverses through a document, he or she is exposed to key concepts/words. First exposure to a given key word is enough to make the reader aware of that subject and that it is relevant to the article<sup>2</sup>. To this end, the Keyword Novelty approach attempts to find when the reader has been exposed to ‘enough’ of these key words that he or she would have a general understanding of the article.

### 4.2.1 Calculating Keyword Novelty

The first step is determining what are the key words in a document. Following standard IR techniques, we limit a document’s text to information-heavy words (nouns), and remove any pluralization through lemmatization. However, not all of the remaining keywords are equally relevant to the document in question. Commonly, term frequency (TF) can be used as a proxy for keyword relevance. However, TF is generally not robust on shorter or sparse data, such as a single newspaper article.

An alternative technique to discovering keyword important is to use Singular Value Decomposition (SVD). SVD is able to filter out the noise in relatively small or sparse data, and is often used for dimensionality reduction. To repurpose SVD to calculate word weight, we represent each sentence as a row in a sentence-word occurrence matrix encompassing  $m$  sentences and  $n$  unique words, which we will refer to as  $\mathbf{M}$  (which can be constructed in  $O(m)$ ).

SVD decomposes the  $m \times n$  matrix  $\mathbf{M}$  into a product of three matrices:  $\mathbf{M} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^*$ .  $\mathbf{\Sigma}$  is a diagonal matrix whose values on the diagonal, referred to as  $\sigma_i$ , are the singular values of  $\mathbf{M}$ . By identifying the four largest  $\sigma_i$  values, which we

<sup>1</sup>It should be noted, that some online websites (e.g. New York Times) create custom one or two sentence previews, rather than drawing them directly from the original text. This research does not examine the quality of these custom hand-crafted previews, which are more akin to single-document summaries.

<sup>2</sup>Exactly how the given subject is related to the full text may not be clear until later in the article.

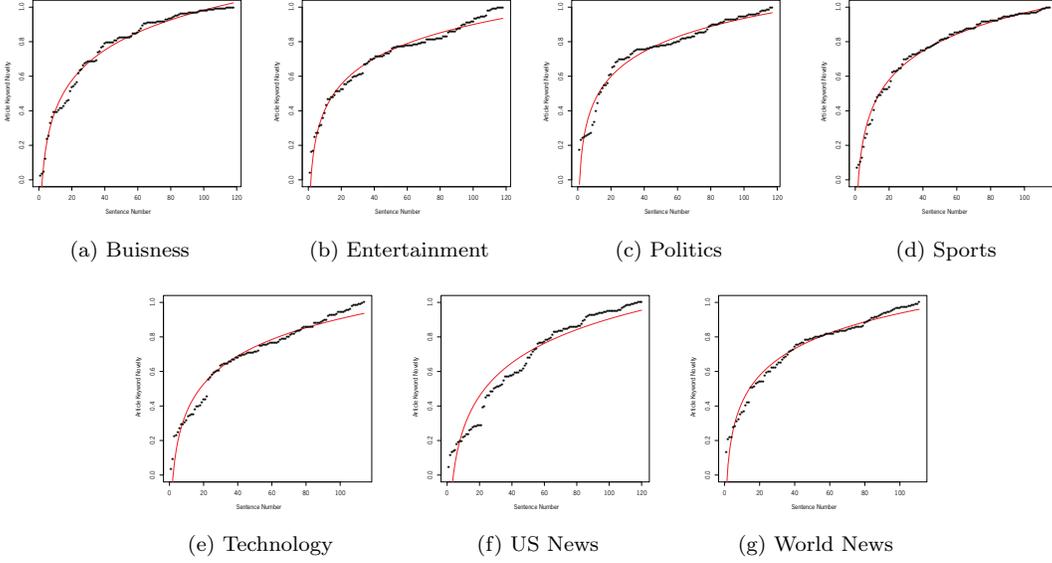


Figure 1: Examples of Article Keyword Novelty Curves

Points are presented with the resulting regression line in red | Graphs are high resolution best viewed in PDF

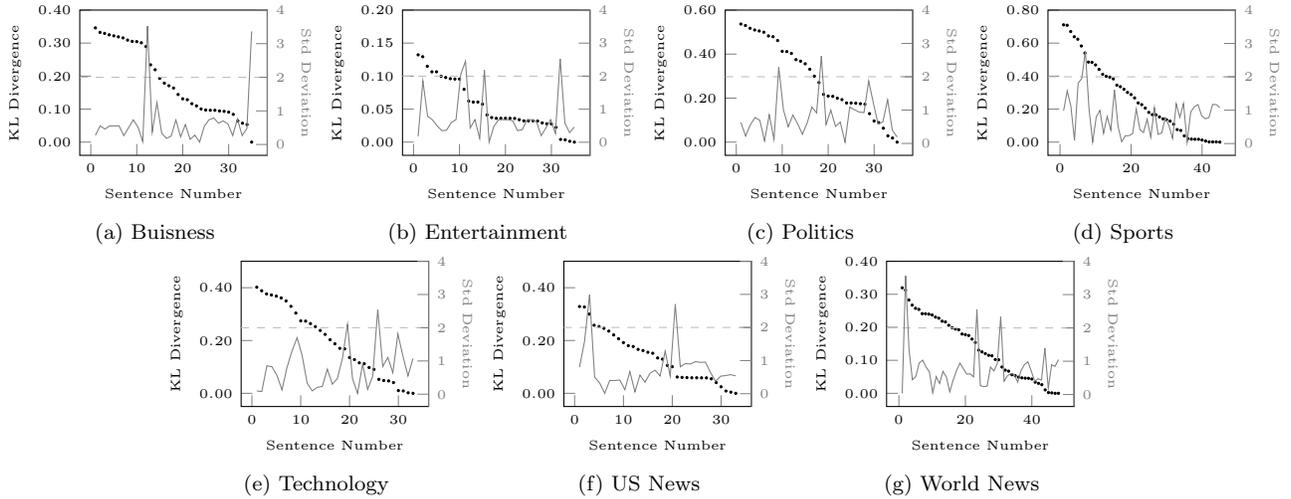


Figure 2: Examples of SLM Article Plots + Std Deviation in Slope Plot

KL Divergence are Black Points (left axis), Standard Deviations Away in Gray (right axis) | Graphs are high resolution best viewed in PDF

refer to as  $\lambda_1 - \lambda_4$ , we are able to take the corresponding top eigenvector columns of  $V$  (which is the conjugate transpose of  $V^*$ ), which we refer to as  $\xi_1 - \xi_4$ . Note that each entry in each of these vectors  $\xi_1 - \xi_4$  corresponds to a unique word in  $M$ .

We then create  $\xi'$ , a master eigenvector calculated as the weighted average of  $\xi_1 - \xi_4$ , weighted by  $\lambda_1 - \lambda_4$ :

$$\xi' = \frac{1}{4} \sum_{i=1}^4 \lambda_i \xi_i \quad (1)$$

Thus,  $\xi'$  is a vector in which each entry represents a unique word, and the value can be interpreted as the ‘centrality’ of the word to the given document [34].

Once we have the keyword weights,<sup>3</sup> we iterate over each sentence in a given article. The ‘value’ of a sentence is the sum of all of the unique keywords’ weights seen up to (and including) that sentence. Thus, the weight of a given sentence is a cumulative sum and each keyword only contributes to the overall sum on its first occurrence.

When plotted, the resulting sequence generally fits a logarithmic curve (see Figure 1), allowing us to consider the inflection point as a ‘lower bound’ break point.<sup>4</sup> The inflection point is an ideal point to paginate in that it is the

<sup>3</sup>To reduce noise, we use only the top 500 words from SVD.

<sup>4</sup>To find the inflection point:

1. Log of the number of sentences (convert to linear)
2. Normalize the weights from 0.0 to 1.0

location where the amount of text needed (space) is increasing more than the amount of new concepts (keywords).

#### 4.2.2 Article vs. Corpus Novelty

Based on the keyword novelty metric, we created two variations:

- **Article Keyword Novelty:** SVD weights are based on the individual article in question
- **Corpus Keyword Novelty:** SVD weights are based on all articles in a corpus treated as a single document

The corpus approach takes a more holistic view of the topic being discussed, whereas the article approach is more sensitive to the specific issues being addressed in a specific article.

### 4.3 Statistical Language Modeling (SLM)

The second approach we present is based on the idea that there is a probabilistic distribution of words (and their frequencies) that are an ‘ideal’ we strive to mimic (e.g. the full text of the article, or the distribution of words in a corpus). We use SLM to create a model of the ‘ideal’ document, and, for a given portion of an article being made visible to the reader, we use an information theoretic measure to discover how closely the model of that article portion comes to the ‘ideal’ model of the entire document.

An added benefit of the SLM approach is the ability to smooth the keyword frequencies that are to common to the broad subject (in this work we use Dirichlet Prior Smoothing, which has been shown to be effective [36]).

#### 4.3.1 Calculating SLM

As in Keyword Novelty, we pre-filter the text in each article to only contain the lemmatized high-information (noun) words.

In order to explain our use of SLM, consider  $\mathbf{S}$  to be the set of all subjects ( $s_0 \dots s_g$ ) in our dataset. We work with one subject at a time, which we will refer to as  $s_i$ .

Let  $\mathbf{D}$  be the set of all articles ( $d_0 \dots d_b$ ) in the given subject  $s_i$  and  $\mathbf{W}$  to be the set of all unique words ( $w_0 \dots w_h$ ) in  $s_i$ .

Denote the frequency of a given word  $w_j$  in a given document  $d_k$  as  $f(w_j|d_k)$ . Then the total count of all words in  $d_k$  is calculated as:

$$T(d_k) = \sum_{j=0}^h f(w_j|d_k) \quad (2)$$

and the probability of a given word ( $w_j$ ) in  $d_k$  is:

$$p(w_j|d_k) = \frac{f(w_j|d_k)}{T(d_k)} \quad (3)$$

This therefore allows us to calculate the probability of a word in a document, using Dirichlet Prior smoothing [36]:

$$q(w_j|d_k) = \frac{f(w_j|d_k) + \mu * p(w_j|s_i)}{T(d_k) + \mu} \quad (4)$$

where  $p(w_j|s_i)$  is the occurrence probability of the word  $w_j$  in the entire subject  $s_i$ :

3. Create a linear regression with the x-axis to be log of the number of sentences, and y-axis to be normalized sentence weights
4. Use the x-axis coefficient (e to the coefficient) as the inflection point

$$p(w_j|s_i) = \frac{\sum_{d_k \in s_i} f(w_j|d_k)}{\sum_{d_k \in s_i} T(d_k)} \quad (5)$$

and where the smoothing constant  $\mu$  is estimated using [31]:

$$m_{w_j} = p(w_j|s_i)$$

$$B_{w_j} = \sum_{d_k \in s_i} \left( \left( \frac{f(w_j|d_k)}{T(d_k)} - m_{w_j} \right)^2 \right) \quad (6)$$

$$\mu = \frac{\sum_{w_j \in W} \frac{B_j}{m_{w_j} * (1 - m_{w_j})}}{\sum_{w_j \in W} \frac{B_j^2}{m_{w_j}^2 * (1 - m_{w_j})^2}}$$

As we traverse an article sentence by sentence, we redefine the variable  $N$  to refer to the subset of sentences in a given article we have seen. Thus, we calculate the probability of a word in  $N$ , using Dirichlet Prior smoothing [36] as:

$$q(w_j|N) = \frac{f(w_j|N) + \mu * p(w_j|s_i)}{T(N) + \mu} \quad (7)$$

To compare each successive test SLM to the Ideal document, we use the KL-Divergence metric (smaller is better):

$$KLDivergence = \sum_{w_j \in d_k} \left( \ln \left( \frac{q(w_j|d_k)}{q(w_j|N)} \right) \right) * q(w_j|d_k) \quad (8)$$

Thus for each sentence in an article, we can calculate the KL Divergence score, returning another set of distributions. When plotted (x-axis=sentence, y-axis=KL Divergence), these distributions appear relatively linear, with occasional ‘jumps’ when the models get closer/further apart (for article and corpus respectively). We can detect the first of these ‘jumps,’ and use it as our break point. To find a jump, we perform the following steps:

1. Calculate the delta in KL-Divergence between having seen  $N$  sentences and  $N + 1$  sentences, for the entire document.
2. Determine the mean and standard deviation of this set of delta values.
3. if the delta in KL-Divergence between having seen  $N$  and  $N + 1$  sentences is greater than or equal to 2 standard deviations away from the mean, there a ‘jump’ after the  $N^{th}$  sentence.<sup>5</sup>

#### 4.3.2 Variations

As with the keyword novelty metric, we created two variations:

- **Article SLM:** The ideal SLM is generated based upon the full text of the specific article
- **Corpus SLM:** The ideal SLM is generated based on all articles in the Corpus treated as a single document

The corpus approach takes a more holistic view of the topic being discussed, whereas the article approach is more sensitive to the specific issues and phrasing being addressed in a specific article.

<sup>5</sup>We ignore the change between the first and second sentence since, due to many articles having ‘low’ value first sentences, having almost any content in the second sentence creates a large delta.

| Subject       | Total #<br>Corpuses | Total #<br>Articles | $\mu$ Articles<br>per Corpus (sd) |
|---------------|---------------------|---------------------|-----------------------------------|
| US            | 183                 | 19429               | 106.17 (52.81)                    |
| Business      | 166                 | 19209               | 115.72 (63.45)                    |
| Politics      | 129                 | 25094               | 194.53 (122.53)                   |
| Entertainment | 218                 | 30534               | 140.06 (78.70)                    |
| World         | 158                 | 20440               | 129.37 (75.59)                    |
| Sports        | 203                 | 27463               | 135.29 (65.08)                    |
| Technology    | 232                 | 42104               | 181.48 (119.72)                   |
| All           | 1289                | 184273              | 142.96 (90.73)                    |

Table 1: Data Collected from Google News from 10/13 to 11/13

## 5. MASTER DATASET: ARTICLE COLLECTION

We collected corpuses of news articles by scraping Google News every 8 hours, beginning on October 21 to November 27 2013<sup>6</sup>. Specifically, we collected corpuses from 7 subjects (World News, US News, Entertainment, Business, Technology, Sports and Politics). Corpuses were ‘accepted’ if they contained at least 50 unique articles. Google News therefore acted as a news aggregator, clustering the articles by topic. From the articles (grouped as a corpus by Google News), we retrieved the source HTML of the original article from the originally hosted website. The body copy of said article was then extracted using CETR [32]. The description of the resulting articles and corpuses can be found in Table 1. To conduct each experiment, we randomly select articles and corpuses from this master dataset.

## 6. EXPERIMENT: MODEL PERFORMANCE

The primary experiment in this paper is to evaluate (in terms of break point semantic content) the performance of the 4 baseline lower bound metrics (One Sentence, Two Sentence, One Paragraph, and Twenty Percent) against the 4 lower bound candidate models (Article Keyword Novelty, Corpus Keyword Novelty, Article SLM, Corpus SLM). In this section we introduce the data set, then describe our methods used for comparing the above 8 algorithms for lower bound break point prediction.

### 6.1 Dataset

100 corpuses from each subject were randomly selected from the master dataset, with one article randomly selected from each corpus. This results in 100 articles from each subject, and 700 articles overall. This allows performance to be compared overall, and take into account the varying writing styles and language within each subject area. Summary statistics including Grade Level<sup>7</sup>, Reading Level<sup>8</sup>, and Fog Index<sup>9</sup> are presented in Table 2.

<sup>6</sup>Collection stopped when at least 100 corpuses per subject area were collected.

<sup>7</sup>Flesch-Kincaid grade level indicates that a student at that current U.S. school grade should be able to understand said document (e.g. 8.0 is eighth grade). 7.0 to 8.0 is ‘optimal.’

<sup>8</sup>The Flesch reading ease rates text on a 100 point scale, with higher scores being easier to understand. 60-70 is ‘optimal.’

<sup>9</sup>Years of education to understand a document in a single reading (e.g. 12.0 is a high school senior). 8.0 is considered ‘optimal.’

## 6.2 Methods

Amazon Mechanical Turk (MT) HITs were constructed from the 700 articles. A HIT is an individual task given to a person on MT. Each HIT consisted of a brief definition of a break point and minimal break point, the original source text of one article<sup>10</sup> and the 8 candidate break points presented in random order (reducing order effects). Before each sentence in the source text, we inserted the characters ‘(SX)’ where X was the sentence number. We used these indexes to refer to a given break point. For each break point, the participants were asked to respond to the statement *Is (SX) a good Minimal Break Point?* with a 7-point scale (*Too Short - Missing text* (1) to *Balanced* (4) to *Too Long - Extra Text* (7)). Each HIT was completed by 2 Master level Turkers<sup>11</sup>, yielding 1400 measures of quality per model (100 articles across 7 areas).

To ensure ‘legitimate’ HIT completion, one ‘sanity check question’ was included asking Turkers to find the Nth word in the Mth sentence. In addition, a HIT was rejected if the Turkers’ response failed basic logic checks on their responses. First, all responses could not be the same. Second, if a single HIT asked about the same cut point more than once, that specific Turker’s response must be the same to both questions (e.g. one Turker rating sentence #4 as a 6 and then rating sentence #4 as a 2 for the same article). Third, ratings must be in chronological order. For example, a sentence early in the document was listed as TOO LONG, and then a sentence later was listed as TOO SHORT, that doesn’t make sense. However succeeding break points (in article order) could have the same ranking. If any of the three logic tests or the sanity check are failed, the HIT is rejected and re-posted. Participants were remunerated 30 cents per accepted HIT.

An ANOVA and Student’s T-test were used to compare the algorithms’ performance. While performing multiple comparisons may suggest statistical adjustment to a more conservative value (i.e., Bonferroni correction), we chose a more transparent method, highlighting multiple thresholds of significance [28]. We further report results and summary statistics broken down by subject area. However, it is outside the scope of this paper to optimize for an individual subject area.

### 6.2.1 Method Limitations

Any evaluation with multiple comparisons has an interaction effect, in that the rating or quality of one break point can be impacted by the other break points offered for comparison. Thus, a break point that might have been rated a 5 (slightly too long) when viewed in isolation, may be pushed ‘higher’ if there are other, less optimal break points located earlier in the article. Thus, these results must necessarily be viewed in the broader context of the other break points presented.

## 6.3 Results

Cross subject results are presented in Table 3 and results by subject area are in Table 4. ANOVA comparing the 8 algorithms resulted in a highly significant difference  $p < 0.001$

<sup>10</sup>HITs were not grouped together (containing more than one article) and presentation order was random (reducing order effects).

<sup>11</sup>HIT Approval Rate above 95% and at least 1000 approved HITS

|                 | Grade Level  | Reading Level | Fog Index    | Sentence Count | Word Count      | Corpus Size     |
|-----------------|--------------|---------------|--------------|----------------|-----------------|-----------------|
| <b>Business</b> | 11.00 (2.68) | 48.54 (8.26)  | 7.33 (2.83)  | 54.52 (23.96)  | 854.57 (151.78) | 112.04 (61.47)  |
| <b>Ent.</b>     | 12.70 (4.18) | 49.60 (13.66) | 10.30 (4.04) | 40.02 (16.27)  | 904 (202.95)    | 148.71 (82.72)  |
| <b>Politics</b> | 13.16 (3.03) | 43.64 (10.77) | 9.72 (3.02)  | 40.15 (13.61)  | 887.93(148.24)  | 187.36 (117.04) |
| <b>Sports</b>   | 11.00 (7.15) | 58.54 (19.28) | 9.55 (7.39)  | 42.11 (13.51)  | 875.98 (176.31) | 127.89 (60.35)  |
| <b>Tech.</b>    | 12.23 (4.37) | 49.55 (14.91) | 9.53 (4.19)  | 42.31 (15.20)  | 903.67 (197.98) | 170.9 (109.85)  |
| <b>US</b>       | 13.14 (3.15) | 45.17 (11.00) | 10.01 (3.12) | 38.83 (12.92)  | 893.36 (190.52) | 109.71 (48.98)  |
| <b>World</b>    | 13.97 (3.13) | 40.32 (11.83) | 10.27 (2.95) | 38.93 (12.57)  | 925.18 (190.72) | 130.38 (72.28)  |
| <b>All</b>      | 12.46 (4.32) | 47.91 (14.23) | 9.53 (4.30)  | 42.41 (16.61)  | 892.10 (184.32) | 141.00 (86.56)  |

Table 2: Documents Source Statistics  
Mean and Std Values Reported

|                          | Algorithm       | Mean (sd)   | Median |
|--------------------------|-----------------|-------------|--------|
| <b>All Subject Areas</b> | First Sentence  | 1.78 (0.97) | 1.00   |
|                          | Second Sentence | 2.47 (1.18) | 2.00   |
|                          | First Paragraph | 2.44 (1.29) | 2.00   |
|                          | Twenty Percent  | 5.19 (1.20) | 5.00   |
|                          | Novelty Article | 5.83 (1.14) | 6.00   |
|                          | Novelty Corpus  | 5.52 (1.19) | 6.00   |
|                          | SLM Article     | 4.66 (1.76) | 5.00   |
|                          | SLM Corpus      | 4.18 (1.65) | 4.00   |

Table 3: Overall Performance  
Ratings based on a 7-Point scale:

(1) “missing information” (4) “balanced” and (7) “excess text.”

(F=1977). Pairwise comparative t-tests between algorithms were likewise all highly significant with  $p < 0.001$ , the one exception being the comparison between Two Sentences and One Paragraph, which had  $p = 0.264$ . This similarity may be due to many articles breaking the first paragraph after two sentences.

Overall, the most balanced approach (minimal extra text while retaining enough key content) was SLM Corpus followed closely by SLM Article. The three static baseline approaches (First Sentence, Second Sentence, First Paragraph) were all overly aggressive, cutting too much text and losing key information that was central to the article themselves. The remaining baseline approach (Twenty Percent) and the two Keyword Novelty methods were generally too relaxed with their break point prediction, choosing to break after superfluous text.

### 6.3.1 Results: Correlations

Given the varying readability and length of the articles (Table 2), we wanted to determine if any of those descriptive features influenced the performance of our algorithms. To this end, we tested the performances of each of the 8 algorithms against all readability statistics found in Table 2 using a Spearman’s rank test. No correlations were detected for any pairing of algorithm and readability statistic (with Spearman’s rho never reaching above 0.2). This suggests that the performances reported in Table 3 & 4 were due to the algorithms themselves, and not the length or readability of the articles.

## 6.4 Discussion: Algorithm Comparison

Overall, the two SLM approaches performed quite well with a mean score of 4.18 and 4.66 for SLM-Corpus and SLM-Article respectively. The improvement in performance between the corpus and article versions is to be expected,

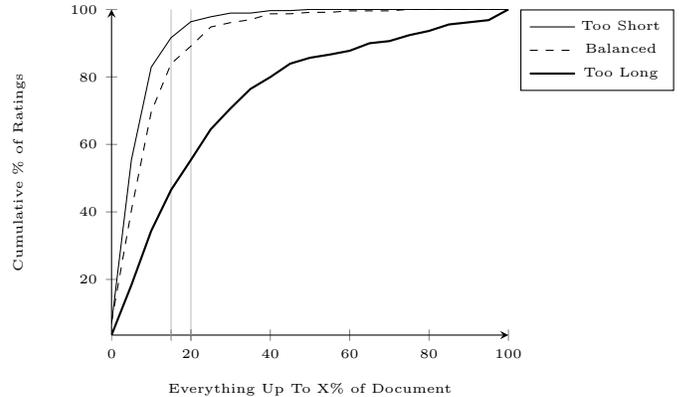


Figure 3: When Failure Occurs  
Gray vertical lines at 15% and 20% for reference

since striving to reach the ideal language model that is the composite of an entire corpus of articles on a topic will smooth out author-specific phrasing, and focus more on the most central key words. However, it is not always realistic to assume that a layout system will have a broader corpus on which to build a more robust language model. Yet SLM Article, which built a language model only from the article, still outperforms the 3 other algorithms with scores above 5 (the two Keyword Novelty approaches, and Twenty Percent).

Given these results, we can also consider First Sentence, Second Sentence, and First Paragraph approaches to be extremely unreasonable for real-world use. When optimizing a document layout, missing key content (score less than 4) is worse than including extra text (score at or above 4), since by leaving out critical information, the reader can be misinformed.

Thus, using SLM objectively produces minimal break points that are more respectful of the semantic content of the article, preventing over-cutting and misleading the readership. In some situations, these break points are more aggressive than other approaches, potentially saving space while maintaining content readability. Yet overall, whether more or less content is needed for a minimal break point, SLM ensures content readability rather than form (layout) over function (readability).

### 6.4.1 A Closer Look at SLM

Based on the high performance of SLM Corpus, we wanted to determine if there was a group of breakpoints that were

|               | Algorithm       | Mean (sd)   | Median |
|---------------|-----------------|-------------|--------|
| Business      | First Sentence  | 2.12 (1.08) | 2.00   |
|               | Second Sentence | 2.83 (1.32) | 3.00   |
|               | First Paragraph | 2.78 (1.41) | 3.00   |
|               | Twenty Percent  | 5.33 (1.23) | 5.00   |
|               | Novelty Article | 5.70 (1.11) | 6.00   |
|               | Novelty Corpus  | 5.61 (1.17) | 6.00   |
|               | SLM Article     | 4.99 (1.81) | 5.00   |
|               | SLM Corpus      | 4.47 (1.64) | 5.00   |
| Entertainment | First Sentence  | 1.43 (0.72) | 1.00   |
|               | Second Sentence | 2.08 (0.92) | 2.00   |
|               | First Paragraph | 2.24 (1.05) | 2.00   |
|               | Twenty Percent  | 5.23 (1.10) | 5.00   |
|               | Novelty Article | 5.88 (1.03) | 6.00   |
|               | Novelty Corpus  | 5.42 (1.05) | 6.00   |
|               | SLM Article     | 4.70 (1.78) | 5.00   |
|               | SLM Corpus      | 3.82 (1.62) | 3.00   |
| Politics      | First Sentence  | 1.72 (0.93) | 1.00   |
|               | Second Sentence | 2.43 (1.17) | 2.00   |
|               | First Paragraph | 2.40 (1.31) | 2.00   |
|               | Twenty Percent  | 5.30 (0.91) | 5.00   |
|               | Novelty Article | 5.89 (0.81) | 6.00   |
|               | Novelty Corpus  | 5.65 (0.94) | 6.00   |
|               | SLM Article     | 4.79 (1.62) | 5.00   |
|               | SLM Corpus      | 4.37 (1.48) | 5.00   |
| Sports        | First Sentence  | 1.62 (0.86) | 1.00   |
|               | Second Sentence | 2.42 (1.06) | 2.00   |
|               | First Paragraph | 2.44 (1.40) | 2.00   |
|               | Twenty Percent  | 5.40 (1.17) | 6.00   |
|               | Novelty Article | 6.12 (1.02) | 6.00   |
|               | Novelty Corpus  | 5.81 (1.10) | 6.00   |
|               | SLM Article     | 4.52 (1.65) | 4.00   |
|               | SLM Corpus      | 4.40 (1.59) | 4.00   |
| Technology    | First Sentence  | 1.82 (0.99) | 1.00   |
|               | Second Sentence | 2.50 (1.31) | 2.00   |
|               | First Paragraph | 2.47 (1.27) | 2.00   |
|               | Twenty Percent  | 5.13 (1.33) | 5.00   |
|               | Novelty Article | 5.78 (1.32) | 6.00   |
|               | Novelty Corpus  | 5.27 (1.42) | 6.00   |
|               | SLM Article     | 4.58 (1.89) | 5.00   |
|               | SLM Corpus      | 4.10 (1.83) | 4.00   |
| US News       | First Sentence  | 1.74 (0.94) | 1.00   |
|               | Second Sentence | 2.39 (1.13) | 2.00   |
|               | First Paragraph | 2.27 (1.27) | 2.00   |
|               | Twenty Percent  | 5.05 (1.17) | 5.00   |
|               | Novelty Article | 5.79 (1.15) | 6.00   |
|               | Novelty Corpus  | 5.50 (1.19) | 6.00   |
|               | SLM Article     | 4.46 (1.81) | 4.00   |
|               | SLM Corpus      | 3.90 (1.61) | 4.00   |
| World News    | First Sentence  | 2.02 (1.07) | 2.00   |
|               | Second Sentence | 2.64 (1.17) | 2.00   |
|               | First Paragraph | 2.49 (1.24) | 2.00   |
|               | Twenty Percent  | 4.88 (1.33) | 5.00   |
|               | Novelty Article | 5.66 (1.39) | 6.00   |
|               | Novelty Corpus  | 5.37 (1.32) | 6.00   |
|               | SLM Article     | 4.55 (1.71) | 5.00   |
|               | SLM Corpus      | 4.20 (1.67) | 4.00   |

Table 4: Performance Per Subject  
Ratings based on a 7-Point scale:

(1) “missing information” (4) “balanced” and (7) “excess text.”

indicative of failure, or success. First, we classified the break point ratings as too short (1-3), balanced (4), and too long (5-7). So as to compare across articles (which have varying length), we divided break points into percentage of document intervals (0-5%, 5-10%, etc), and tabulated the number of ratings in each interval for a given class. To compare across the three classes, we must consider the tabulation as a percentage of the total ratings in said class. The results are plotted in Figure 3.

Upon viewing the curves for the three classes, it is worth noting that the ‘Too Short’ and ‘Balanced’ curves are almost identical. More specifically, 82% of the failed, 69% of the balanced, and 34% of the too long occur at less than 15% of the document. These results suggest why absolute approaches (e.g. Twenty percent) may not be aggressive enough. Further, these results also indicate that there does not appear to be a universal lower threshold below which we are assured to be ‘too short.’ Thus, having adaptive algorithms that are dependent upon the semantic content of a document is critical to achieving quality pagination.

## 7. GOLD STANDARD GENERATION

While the above work strove to identify an improved algorithm to better predict a lower bound break point based on semantics, we postulated that there may be a truly ‘ideal’ break point in each article. This ‘ideal’ would be the natural place to break the article, and unlike the minimal break point, this ideal could be longer and contain more than just the bare minimum of information. To this end, we attempted to create a ‘gold standard’ dataset.

### 7.1 Study Design

The goal of this study was to find consensus on an ideal break point for a given article for the construction of a gold standard. We followed the same random selection technique as described in Section 6, pulling 700 unique articles (100 from each subject area). We used CrowdFlower (CF) as our crowd-sourced platform. Unlike MT, CF has a premium set of crowd-sourced individuals called the Editorial Crowd. From the CF website:

*This group of contributors have been tested for a deep understanding of the English language. These contributors have shown that they understand syntax, grammar, punctuation, and other elements of the English language.*

We realize that this task, attempting to find consensus, is a challenge and we therefore opted to use the more expensive CF verified workers with an expertise in English editing to ensure a careful reading and quality consideration of the ideal break point. We will refer to workers on CF as **CFWs**.

CF HITs were constructed from the 700 randomly selected articles. A HIT consisted of a brief definition of a break point and the original source text (using the same sentence numbering as in the above experiment). The text was accompanied by the question, ‘What is the best break point in this document?’ CFWs were asked to choose one of the sentence numbers as the best break point (indicating everything before that number would be kept, and everything after would be cut). A full definition of an “ideal” break point was included to clear up any ambiguity on the labeling task including the statement “*This should be the shortest document possible without sacrificing key content.*” To ensure legitimate HIT completion, one ‘sanity check question’

| Sentence Tolerance | No CFW Agree | 2 CFW Agree | 3 CFW Agree | 4 CFW Agree | 5 CFW Agree |
|--------------------|--------------|-------------|-------------|-------------|-------------|
| 0 Away             | 361          | 292         | 46          | 1           | 0           |
| 1 Away             | 168          | 345         | 139         | 34          | 15          |
| 2 Away             | 92           | 304         | 201         | 68          | 35          |
| 3 Away             | 42           | 263         | 240         | 95          | 60          |

Table 5: Gold Standard Consensus for 700 Random Articles  
*Values represent the number of articles where consensus was at the column level given the row’s tolerance. For example, there were 139 articles where 3 CFWs agreed on a break point within a 1 sentence tolerance (those three breakpoints were separated by no more than one sentence.)*

was included asking CFWs to find the Nth word in the Mth sentence. Failed sanity check question HITS were reposted.

The CF infrastructure groups HITS together. Thus, each CFW was required to complete 4 HITS at a time, and was remunerated a total of 40 cents for those 4 HITS. Every HIT was evaluated by 5 unique CFWs, totaling 3500 observations.

Given the subjective nature of these observations, we wished to only accept a break point (and its corresponding article) as a gold standard if a majority (at least 3 of the 5 CFWs) could agree upon an ideal break point, within a tolerance.

## 7.2 Results & Discussion

In a surprising turn of events, we saw an extremely low level of CFW agreement, as illustrated in Table 5. As part of our analysis, we include a range of agreement tolerance values (how many sentences are allowed to separate the break points chosen by k CFWs such that they could still be said to agree?) With the strictest tolerance (break points must exactly match, only 46 of the 700 articles had 3 out of 5 CFWs agreeing, 1 had 4 CFWs agreeing, and no articles had all 5 CFWs agreeing. This results in only 6.7% of the articles reaching any semblance of consensus. Even at a very generous 3 sentence tolerance (there may be up to 3 sentences between break points chosen by k CFWs, and agreement could still be claimed), only 395 articles (slightly above 50%) reached a consensus at or above 3 CFWs.

Given these low consensus results, the ability to meaningfully construct a gold standard is extremely limited. This suggests that the use of machine learning techniques, which rely upon a training set, would be prohibitive in this context.

Furthermore, this lack of consensus may suggest that there is no single ‘ideal’ break point in a given article. Beyond finding a minimal ‘do not cut before this place’ break point, the remaining text of the article (and its importance to the reader) may be too subjective to readily predict, and subsequently, this task may be “too hard for a human.” This adds further support to the importance and tractable outcome observed in our minimal break point detection.

## 8. FUTURE WORK

The first area of future work would be to develop a variation of PDM [8, 1] that integrates pagination quality as part of the layout quality metric. Currently these approaches are based on fit of content alone. Subsequently, integrating pagination quality would be a substantial challenge, but a

critical next step to testing the impact of pagination break point quality in an existing print publications.

Second, we examine the applications of pagination break points’ application to e-documents. In this context, there is an explicit metric of quality to test layout impact: number of clicks of a “read more” link after the initial document content. This is another exciting application that would require a large amount of structural work to set up an experiment.

One last area of future work is to explore alternative statistical language models, including those that do not accord equal weight to each occurrence of a word. This could further improve SLM performance, especially for the SLM Article context. We still believe that the Keyword Novelty approach has merit; however a more ‘aggressive’ keyword weighting algorithm may be needed to cause the curves’ inflection point to occur earlier in the article.

## 9. CONCLUSION

Automatic document layout is rapidly becoming central to the production of most commercially printed newspapers and magazines. Within this context, pagination is a common layout challenge. However, all existing approaches to automatically calculate article break points for pagination neglect to account for the semantic content of the presented article. Disregarding the semantic information within the document can lead to overly aggressive or overly relaxed cutting, thereby running the risk of either eliminating key content and confusing the reader, or leaving in superfluous content and boring the reader, as well as making automatic layout more difficult.

We seek to directly address this shortcoming in this work. We present the first semantic-based pagination algorithm, to the authors’ knowledge, for news article layout. Our approach predicts minimal break points based on the semantic content of an article through the use of a statistical language model. This approach is tested via a multi-subject experiment on 700 documents, comparing our method to 4 currently employed baselines and 3 alternative semantic approaches (also created for this paper). Results from this experiment clearly show that our approach strongly outperforms the baselines and alternatives. To further explore break point detection for pagination, we investigate whether an ‘ideal’ break point could be found for paginating news articles. Results from a second study suggest that humans are not able to agree on a single best break point, suggesting that it is more practical to define reliable minimal break points for pagination tasks.

This work presents a strong validation that break point detection for pagination tasks can benefit from an examination of the semantic content of the news articles themselves. Within a real-world context, a combination of semantic-based lower bound break point prediction and spacial/aesthetic optimization is ideal for automated document synthesis.

## 10. REFERENCES

- [1] I. Ahmadullin and N. Damera-Venkata. Hierarchical probabilistic model for news composition. In *DocEng*, page 141, New York, New York, USA, Sept. 2013. ACM Request Permissions.
- [2] G. J. Badros, A. Borning, and P. J. Stuckey. The Cassowary linear arithmetic constraint solving algorithm. *TOCHI*, 8(4 (Dec)):267–306, Dec. 2001.

- [3] D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. *Machine learning*, 34(1-3):177–210, 1999.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.
- [5] A. Brüggemann-Klein, R. Klein, and S. Wohlfeil. On the pagination of complex documents. *Lecture Notes in Computer Science*, 2598:49–68, 2003.
- [6] F. Chua and S. Asur. Automatic Summarization of Events From Social Media. In *ICWSM*, 2013.
- [7] P. Ciancarini, A. Di Iorio, L. Furini, and F. Vitali. High-quality pagination for publishing. *Software—Practice & Experience*, 42(6), June 2012.
- [8] N. Damera-Venkata, J. Bento, and E. O’Brien-Strain. Probabilistic document model for automated document composition. In *DocEng*, page 3, New York, New York, USA, Sept. 2011. ACM Request Permissions.
- [9] H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the ACM (JACM)*, 16(2), Apr. 1969.
- [10] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *J Artif Intell Res(JAIR)*, 2004.
- [11] M. Fiszman, T. C. Rindfleisch, and H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. pages 76–83, May 2004.
- [12] F. Giannetti. An exploratory mapping strategy for web-driven magazines. In *Proceeding of the eighth ACM symposium*, pages 223–229, New York, New York, USA, 2008. ACM Press.
- [13] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. pages 362–370, May 2009.
- [14] J. Hailpern, N. Damera Venkata, and M. Danilevsky. Truncation: All the News that Fits WeãŽll Print. In *DocENG*. ACM, 2014.
- [15] J. Hailpern and B. A. Huberman. Echo: the editor’s wisdom with the elegance of a magazine. In *EICS*. ACM Request Permissions, June 2013.
- [16] M. A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1), Mar. 1997.
- [17] N. Hurst, W. Li, and K. Marriott. Review of automatic document formatting. In *DocEng*, page 99, New York, New York, USA, Sept. 2009. ACM Request Permissions.
- [18] C. Jacobs, W. Li, E. Schrier, D. Barger, and D. Salesin. Adaptive grid-based document layout. *SIGGRAPH*, 22(3):838–847, July 2003.
- [19] N. Jamil, J. Mueller, C. Lutteroth, and G. Weber. Extending Linear Relaxation for User Interface Layout. In *ICTAI*. IEEE Computer Society, Nov. 2012.
- [20] I. Kastner and C. Monz. Automatic single-document key fact extraction from newswire articles. In *EACL*. Association for Computational Linguistics, Mar. 2009.
- [21] R. Katragadda, P. Pingali, and V. Varma. Sentence position revisited: a robust light-weight update summarization baseline algorithm. pages 46–52, June 2009.
- [22] C.-Y. Lin and E. Hovy. Identifying topics by position. In *ANCL*. Association for Computational Linguistics, Mar. 1997.
- [23] C. Lutteroth, R. Strandh, and G. Weber. Domain Specific High-Level Constraints for User Interface Layout. *Constraints*, 13(3), Sept. 2008.
- [24] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *EMNLP*, 2004.
- [25] A. Nenkova. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *AAAI*. AAAI Press, July 2005.
- [26] A. Nenkova and L. Vanderwende. The impact of frequency on summarization. Technical Report MSR-TR-2005-101, Microsoft Research, 2005.
- [27] L. Pevzner and M. A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 2002.
- [28] D. A. Savitz and A. F. Olshan. Multiple comparisons and related issues in the interpretation of epidemiologic data. 1995.
- [29] A. Scoditti and W. Stuerzlinger. A new layout method for graphical user interfaces. In *TIC-STH*, pages 642–647. IEEE, 2009.
- [30] Y. Seki, K. Eguchi, and N. Kando. Compact Summarization for Mobile Phones. *Mobile and Ubiquitous Information Access*, 2954(Chapter 13):172–186, 2004.
- [31] J. Seo and W. B. Croft. Unsupervised estimation of dirichlet smoothing parameters. In *SIGIR ’10*, pages 759–760, New York, New York, USA, 2010. ACM Press.
- [32] T. Wenginger, W. H. Hsu, and J. Han. CETR: content extraction via tag ratios. *WWW 2010*, 2010.
- [33] C. C. Yang and F. L. Wang. Automatic summarization of financial news delivery on mobile devices. In *WWW’03*, 2003.
- [34] C. C. Yang and F. L. Wang. Hierarchical summarization of large documents. *J. of the American Society for Information Science and Technology*, 59(6), Apr. 2008.
- [35] C. Zeidler, J. Müller, C. Lutteroth, and G. Weber. Comparing the usability of grid-bag and constraint-based layouts. In *OzCHI*, pages 674–682, New York, New York, USA, Nov. 2012. ACM Request Permissions.
- [36] C. Zhai. *Statistical Language Models for Information Retrieval*. Morgan & Claypool Publishers, 2009.