5-2013

New Alignment Methods for Discriminative Book Summarization

David Bamman
Carnegie Mellon University

Noah A. Smith
Carnegie Mellon University, nasmith@cs.cmu.edu

New Alignment Methods for Discriminative Book Summarization

Work in Progress

David Bamman and **Noah A. Smith**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dbamman,nasmith}@cs.cmu.edu

Abstract

We consider the unsupervised alignment of the full text of a book with a human-written summary. This presents challenges not seen in other text alignment problems, including a disparity in length and, consequent to this, a violation of the expectation that individual words and phrases *should* align, since large passages and chapters can be distilled into a single summary phrase. We present two new methods, based on hidden Markov models, specifically targeted to this problem, and demonstrate gains on an extractive book summarization task. While there is still much room for improvement, unsupervised alignment holds intrinsic value in offering insight into what features of a book are deemed worthy of summarization.

1 Introduction

The task of *extractive summarization* is to select a subset of sentences from a source document to present as a summary. Supervised approaches to this problem make use of training data in the form of source documents paired with existing summaries (Marcu, 1999; Osborne, 2002; Jing and McKeown, 1999; Ceylan and Mihalcea, 2009). These methods learn what features of a source sentence are likely to result in that sentence appearing in the summary; for news articles, for example, strong predictive features include the position of a sentence in a document (earlier is better), the sentence length (shorter is better), and the number of words in a sentence that are among the most frequent in the document.

Supervised discriminative summarization relies on an alignment between a source document and its summary. For short texts and training pairs where a one-to-one alignment between source and abstract sentences can be expected, standard techniques from machine translation can be applied, including word-level alignment (Brown et al., 1990; Vogel et al., 1996; Och and Ney, 2003) and longer phrasal alignment (Daumé and Marcu, 2005), especially as adapted to the monolingual setting (Quirk et al., 2004). For longer texts where inference over all possible word alignments becomes intractable, effective approximations can be made, such as restricting the space of the available target alignments to only those that match the identity of the source word (Jing and McKeown, 1999).

The use of alignment techniques for book summarization, however, challenges some of these assumptions. The first is the disparity between the length of the source document and that of a summary. While the ratio between abstracts and source documents in the benchmark Ziff-Davis corpus of newswire (Marcu, 1999) is approximately 12% (133 words vs. 1,066 words), the length of a full-text book greatly overshadows the length of a simple summary. Figure 1 illustrates this with a dataset comprised of books from Project Gutenberg paired with plot summaries extracted from Wikipedia for a set of 439 books (described more fully in §4.1 below). The average ratio between a summary and its corresponding book is 1.2%.

This disparity in size leads to a potential violation of a second assumption: that we expect words and phrases in the source document to align with words and phrases in the target. When the disparity is so great, we might rather expect that an entire paragraph, page, or even chapter in a book aligns to a single summary sentence.

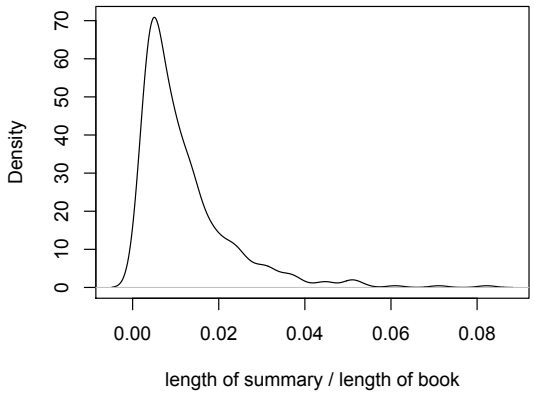

Figure 1: Size disparity between summaries and full texts. Summaries average 1% the size of the corresponding book. The mean is 0.012, with a $[5, 95]$ quantile of $[0.002, 0.032]$.

To help adapt existing methods of supervised document summarization to books, we present two alignment techniques that are specifically adapted to the problem of book alignment, one that aligns passages of varying size in the source document to sentences in the summary, guided by the unigram language model probability of the sentence under that passage; and one that generalizes the HMM alignment model of Och and Ney (2003) to the case of long but sparsely aligned documents.

2 Related Work

This work builds on a long history of unsupervised word and phrase alignment originating in the machine translation literature, both for the task of learning alignments across parallel text (Brown et al., 1990; Vogel et al., 1996; Och and Ney, 2003; DeNero et al., 2008) and between monolingual (Quirk et al., 2004) and comparable corpora (Barzilay and Elhadad, 2003). For the related task of document/abstract alignment, we draw on work in document summarization (Marcu, 1999; Osborne, 2002; Daumé and Marcu, 2005). Past approaches to fictional summarization, including both short stories (Kazantseva and Szpakowicz, 2010) and books (Mihalcea and Ceylan, 2007), have tended toward non-discriminative methods; one notable exception is Ceylan (2011), which applies the Viterbi alignment method of Jing and McKeown (1999) to a set of 31 literary novels.

3 Methods

We present two methods, both of which involve estimating the parameters of a hidden Markov model (HMM). The HMMs differ in their definitions of states, observations, and parameterizations of the emission distributions. We present a generic HMM first, then instantiate it with each of our two models, discussing their respective inference and learning algorithms in turn.

Let $\mathcal{S}$ be the set of hidden states and $K = |\mathcal{S}|$. An observation sequence $\boldsymbol{t} = \langle t_1, \ldots, t_n \rangle$, each $t_\ell \in \mathcal{V}$, is assigned probability:

$$p(\boldsymbol{t} \mid n) = \sum_{\boldsymbol{z} \in \mathcal{S}^n} \pi_{z_1} \left(\prod_{\ell=1}^{n} \eta_{z_\ell, t_\ell} \gamma_{z_\ell, z_{\ell+1}} \right) \quad (1)$$

where $\boldsymbol{z}$ is the sequence of hidden states, $\boldsymbol{\pi} \in \Delta_K$ is the distribution over start states, and for all $s \in \mathcal{S}$, $\boldsymbol{\eta}_s \in \Delta_{|\mathcal{V}|}$ and $\boldsymbol{\gamma}_s \in \Delta_K$ are s's emission and transition distributions, respectively. Note that we avoid stopping probabilities by always conditioning on the sequence length.

3.1 Passage Model

In the passage model, each HMM state corresponds to a contiguous passage in the source document. The intuition behind this approach is the following: while word and phrasal alignment attempts to capture fine-grained correspondences between a source and target document, longer documents that are distilled into comparatively short summaries may instead have long, topically coherent passages that are summarized into a single sentence. For example, the following summary sentence in a Wikipedia plot synopsis summarizes several long episodic passages in *The Adventures of Tom Sawyer*:

> After playing hooky from school on Friday and dirtying his clothes in a fight, Tom is made to whitewash the fence as punishment all of the next day.

Our aim is to find the sequence of passages in the source document that aligns to the sequence of summary sentences. Therefore, we identify each HMM

	Passage model	Token model
states $\mathcal{S}$	source document passages	source document tokens
observations	summary sentences	summary tokens
transitions	by passage order difference	by distance bin
emissions	unigram distribution	lexical identity, synonyms

Table 1: Summary of the passage model (§3.1) and the token model (§3.2).

state in $s \in \mathcal{S}$ with source document positions i_s and j_s. When a summary sentence $t_\ell = \langle t_{\ell,1}, \ldots, t_{\ell,T_\ell} \rangle$ is sampled from state s, its emission probability is defined as follows:

$$\eta_{s,t_\ell} = \prod_{k=1}^{T_\ell} \hat{p}_{unigram}(t_{\ell,k} \mid \boldsymbol{b}_{i_s:j_s}) \qquad (2)$$

where $\boldsymbol{b}_{i_s:j_s}$ is the passage in the source document from position i_s to position j_s; again, we avoid a stop symbol by implicitly assuming lengths are fixed exogenously. The unigram distribution $\hat{p}_{unigram}(\cdot \mid \boldsymbol{b}_{i_s:j_s})$ is estimated directly from the source document passage $\boldsymbol{b}_{i_s:j_s}$.

The transition distribution from state $s \in \mathcal{S}$, $\boldsymbol{\gamma}_s$ is operationalized following the HMM word alignment formulation of Vogel et al. (1996). The transition events between ordered pairs of states are binned by the difference in two passages' ranks within the source document.[1] We give the formula for relative frequency estimation of the transition distributions:

$$\gamma_{s,s'} = \frac{c(s' - s)}{\sum_{s'' \in \mathcal{S}} c(s - s'')} \qquad (3)$$

where $c(\cdot)$ denotes the count of jumps of a particular length, measured as the distance between the rank order of two passages within a document; the count of a jump between passage 10 and passage 13 is the same as that between passage 21 and 24; namely, $c(3)$. Note that this distance is signed, so that the distance of a backwards jump from passage 13 to passage 10 (-3) is not the same as a jump from 10 to 13 (3).

The HMM states' spans are constrained not to overlap with each other, and they need not cover the source document. Because we do not know the boundary positions for states in advance, we must estimate them alongside the traditional HMM parameters. Figure 2 illustrates this scenario with a sequence of 17 words in the source document ($[1 \ldots 17]$) and 4 sentences in the target summary ($\{a, b, c, d\}$). In this case, the states correspond to $[1 \ldots 4]$, $[9 \ldots 13]$, and $[15 \ldots 17]$.

[1]These ranks are fixed; our inference procedure does not allow passages to overlap or to "leapfrog" over each other across iterations.

3.1.1 Inference

Given a source document $\boldsymbol{b}$ and a target summary $\boldsymbol{t}$, our aim is to infer the most likely passage z_ℓ for each sentence t_ℓ. This depends on the parameters ($\boldsymbol{\pi}$, $\boldsymbol{\eta}$, and $\boldsymbol{\gamma}$) and the passages associated with each state, so we estimate those as well, seeking to maximize likelihood. Our approach is an EM-like algorithm (Dempster et al., 1977); after initialization, it iterates among three steps:

- *E-step.* Calculate $p(\boldsymbol{t})$ and the posterior distributions $q(z_k \mid \boldsymbol{t})$ for each sentence t_k. This is done using the forward-backward algorithm.
- *M-step.* Estimate $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ from the posteriors, using the usual HMM M-step.
- *S-step.* Sample new passages for each state. The sampling distribution considers, for each state s, moving i_s subject to the no-overlapping constraint and j_s, and then moving j_s subject to the no-overlapping constraint and i_s (DeNero et al., 2008). (See §3.1.2 for more details.) The emission distribution $\boldsymbol{\eta}_s$ is updated whenever i_s and j_s change, through Equation 2.

For the experiments described in section 4, each source document is initially divided into K equal-length passages ($K = 100$), from which initial emission probabilities are defined; $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ are both initialized to uniform distribution. Boundary samples are collected once for each iteration, after one E step and one M step, for a total of 500 iterations.

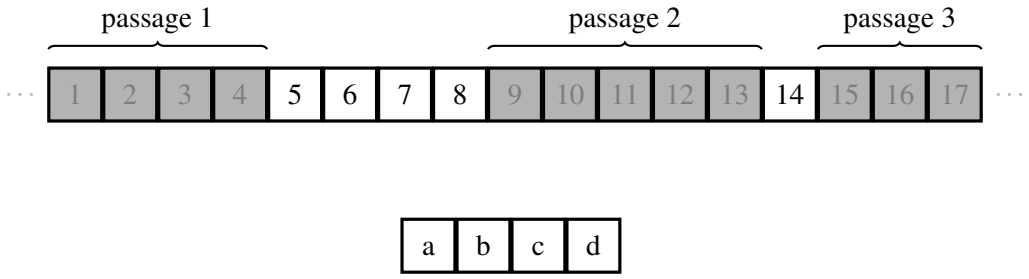

Figure 2: Illustration of the passage HMM. HMM states correspond to passages in the source document (top); each emission is a summary sentence (bottom).

3.1.2 Sampling chunk boundaries

During the S-step, we sample the boundaries of each HMM state's passage, favoring (stochastically) those boundaries that make the observations more likely. We expect that, early on, most chunks will be radically reduced to smaller spans that match closely the target sentences aligned to them with high probability. Over subsequent iterations, longer spans should be favored when adding words at a boundary offsets the cost of adding the non-essential words between the old and new boundary.

A greedy step—analogous to the M-step use to estimate parameters—is one way to do this: we could, on each S-step, move each span's boundaries to the positions that maximize likelihood under the revised language model. Good local choices, however, may lead to suboptimal global results, so we turn instead to sampling. Note that, if our model defined a marginal distribution over passage boundary positions in the source document, this sampling step could be interpreted as part of a Markov Chain Monte Carlo EM algorithm (Wei and Tanner, 1990). As it is, we do not have such a distribution; this equates to a fixed uniform distribution over all valid (non-overlapping) passage boundaries.

The implication is that the probability of a particular state s's passage's start- or end-position is proportional to the probability of the observations generated given that span. Following any E-step, the assignment of observations to s will be fractional. This means that the likelihood, as a function of particular values of i_s and j_s, depends on all of the sentences in the summary:

$$L(i_s, j_s) = \prod_{\ell=1}^{n} \eta_{s,t_\ell}^{q(z_\ell=s|\boldsymbol{t})} \tag{4}$$
$$= \prod_{\ell=1}^{n} \left(\prod_{k=1}^{T_\ell} \hat{p}_{unigram}(t_{\ell,k} \mid \boldsymbol{b}_{i_s:j_s}) \right)^{q(z_\ell=s|\boldsymbol{t})}$$

For example, in Figure 2, the start position of the second span (word 9) might move anywhere from word 5 (just past the end of the previous span) to word 12 (just before the end of its own span, $j_s = 12$). Each of the values should be sampled with probability proportional to Equation 4, so that the sampling distribution is:

$$\frac{1}{\sum_{i=5}^{12} L(i, 12)} \langle L(5, 12), L(6, 12), \ldots, L(12, 12) \rangle$$

Calculating L for different boundaries requires recalculating the emission probabilities η_{s,t_ℓ} as the language model changes. We can do this efficiently (in linear time) by decomposing the language model probability. Here we represent a state s by its boundary positions in the source document, $i : j$, and we use the relative frequency estimate for $\hat{p}_{unigram}$.

$$\log \eta_{i:j,t_\ell} = \sum_{k=1}^{T_\ell} \log \frac{\mathit{freq}(t_{\ell,k}; \boldsymbol{b}_{i:j})}{j - i + 1} \tag{5}$$
$$= -T_\ell \log(j - i + 1) + \sum_{k=1}^{T_\ell} \log \mathit{freq}(t_{\ell,k}; \boldsymbol{b}_{i:j}) \tag{6}$$

Now consider the change if we remove the first word from s's passage, so that its boundaries are $[i+1, j]$.

Let b_i denote the source document's word at position i. $\log \eta_{i+1:j,t_\ell} =$

$$-T_\ell \log(j-i) + \sum_{k=1}^{T_\ell} \log freq(t_{\ell,k}; \boldsymbol{b}_{i+1:j})$$
$$= \log \eta_{i:j,t_\ell} + freq(b_i; t_\ell) \log \frac{freq(b_i; \boldsymbol{b}_{i:j}) - 1}{freq(b_i; \boldsymbol{b}_{i:j})}$$
$$+ T_\ell \log \frac{j-i+1}{j-i} \quad (7)$$

This recurrence is easy to solve for all possible left boundaries (respecting the no-overlap constraints) if we keep track of the word frequencies in each span of the source document—something we must do anyway to calculate $\hat{p}_{unigram}$. A similar recurrence holds for the right boundary of a passage.

Figure 3 illustrates the result of this sampling procedure on the start and end positions for a single source passage in *Heart of Darkness*. After 500 iterations, the samples can be seen to fluctuate over a span of approximately 600 words; however, the modes are relatively peaked, with the most likely start position at 1613, and the most likely end position at 1660 (yielding a span of 47 words).

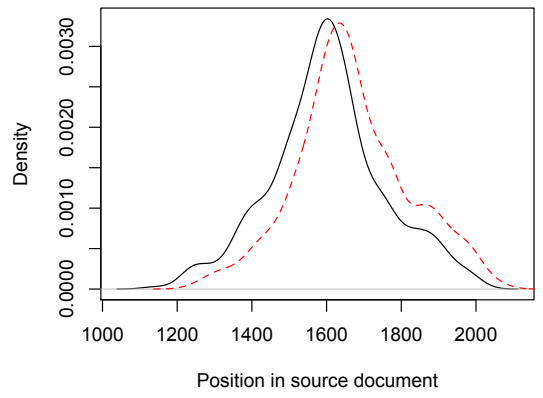

Figure 3: Density plot of accumulated samples for one passage HMM state, in *Heart of Darkness*. The left boundary is shown in black and solid, the right boundary in red and dashed.

3.2 Token Model

Jing and McKeown (1999) introduced an HMM whose states correspond to tokens in the source document. The observation is the sequence of target summary tokens (restricting to those types found in the source document). The emission probabilities are fixed to be one if the source and target words match, zero if they do not. Hence each instance of $v \in \mathcal{V}$ in the target summary is assumed to be aligned to an instance of v in the source. The transition parameters were fixed manually to simulate a ranked set of transition types (e.g., transitions within the same sentence are more likely than transitions between sentences). No parameter estimation is used; the Viterbi algorithm is used to find the most probable alignment. The allowable transition space is bounded by F^2, where F is the frequency of the most common token in the source document. The resulting model is scalable to large source documents (Ceylan and Mihalcea, 2009; Ceylan, 2011).

One potential issue with this model is that it lacks the concept of a null source, not articulated in the original HMM alignment model of Vogel et al. (1996) but added by Och and Ney (2003). Without such a null source, every word in the summary must be generated by some word in the source document. The consequence of this decision is that a Viterbi alignment over the summary must pick a perhaps distant, low-probability word in the source document if no closer word is available. Additionally, while the choice to enforce lexical identity constrains the state space, it also limits the range of lexical variation captured.

Our second model extends Jing's approach in three ways.

First, we introduce parameter inference to learn the values of start probabilities and transitions that maximize the likelihood of the data, using the EM algorithm. We operationalize the transition probabilities again following Vogel et al. (1996), but constrain the state space by only measuring transitiitions between fixed bucket lengths, rather than between the absolute position of each source word. The relative frequency estimator for transitions is:

$$\gamma_{s,s'} = \frac{c(b(s'-s))}{\sum_{s'' \in \mathcal{S}} c(b(s''-s))} \quad (8)$$

As above, $c(\cdot)$ denotes the count of an event, and here $b(\cdot)$ is a function that transforms the difference between two token positions into a coarser set of bins (for example, b may transform a distance of 0

into its own bin, a distance of $+1$ into a different bin, a distance in the range $[+2, +10]$ into a third bin, a difference of $[-10, -2]$ into a fourth, etc.). Future work may include dynamically learning optimizal bin sizes, much as boundaries are learned in the passage HMM.

Second, we introduce the concept of a null source that can generate words in the target sentence. In the sentence-to-sentence translation setting, for a source sentence that is m words long, Och and Ney (2003) add m corresponding NULL tokens, one for each source word position, to be able to adequately model transitions to, from and between NULL tokens in an alignment. For a source *document* that is ca. 100,000 words long, this is clearly infeasible (since the complexity of even a single round of forward-backward inference is $O(m^2 n)$, where n is the number of words in the target summary $\boldsymbol{t}$). However, we can solve this problem by noting that the transition probability as defined above is not measured between individual words, but rather between the positions of coarser-grained chunks that contain each word; by coarsing the transitions to model the jump between a fixed set of B bins (where $B \ll m$), we effectively only need to add B null tokens, making inference tractable. As a final restriction, we disallow transitions between source state positions i and j where $|i - j| > \tau$. In the experiments described in section 4, $\tau = 1000$.

Third, we expand the emission probabilities to allow the translation of a source word into a fixed set of synonyms (e.g., as derived from Roget's Thesaurus.[2]) This expands the coverage of important lexical variants while still constraining the allowable emission space to a reasonable size. All synonyms of a word are available as potential "translations"; the exact translation probability (e.g., $\eta_{\text{purchase,buy}}$) is learned during inference.

4 Experiments

To evaluate these two alignment methods and compare with past work, we evaluate on the downstream task of extractive book summarization.

[2] http://www.gutenberg.org/ebooks/10681

4.1 Data

The available data includes 14,120 book plot summaries extracted from the November 2, 2012 dump of English-language Wikipedia[3] and 31,393 English-language books from Project Gutenberg.[4]. We restrict the book/summary pairs to only those where the full text of the book contains at least 10,000 words and the paired abstract contains at least 100 words (stopwords and punctuation excluded). This results in a dataset of 439 book/summary pairs, where the average book length is 43,223 words, and the average summary length is 369 words (again, not counting stopwords and punctuation).

The ratio between summaries and full books in this dataset is approximately 1.2%, much smaller than that used in previous work for any domain, even for past work involving literary novels: Ceylan (2009) makes use of a collection of 31 books paired with relatively long summaries from SparkNotes, CliffsNotes and GradeSaver, where the average summary length is 6,800 words. We focus instead on the more concise case, targeting summaries that distill an entire book into approximately 500 words.

4.2 Discriminative summarization

We follow a standard approach to discriminative summarization. All experiments described below use 10-fold cross validation, in which we partition the data into ten disjoint sets, train on nine of them and then test on the remaining held-out partition. Ten evaluations are conducted in total, with the reported accuracy being the average across all ten sets. First, all source books and paired summaries in the training set are aligned using one of the three unsupervised methods described above (Passage HMM, Token HMM, Jing 1999).

Next, all of the sentences in the source side of the book/summary pairs are featurized; all sentences that have been aligned to a sentence in the summary are assiged a label of 1 (appearing in summary) and 0 otherwise (not appearing in summary). Using this featurized representation, we then train a binary logistic regression classifier with ℓ_2 regularization on the training data to learn which features are the most

[3] http://dumps.wikimedia.org/enwiki/

[4] http://www.gutenberg.org

indicative of a source sentence appearing in a summary. Following previous work, we devise sentence-level features that can be readily computed in comparison both with the document in which the sentence in found, and in comparison with the collection of documents as whole (Yeh et al., 2005; Shen et al., 2007). All feature values are binary:

- Sentence position within document, discretized into membership in each of ten deciles. (10 features.)
- Sentence contains a salient name. We operationalize "salient name" as the 100 capitalized words in a document with the highest TF-IDF score in comparison with the rest of the data; only non-sentence-initial tokens are used for calculate counts. (100 features.)
- Contains lexical item x ($x \in$ most frequent 10,000 words). This captures the tendency for some actions, such as *kills, dies* to be more likely to appear in a summary. (10,000 features.)
- Contains the *first* mention of lexical item x ($x \in$ most frequent 10,000 words). (10,000 features.)
- Contains a word that is among the top [1,10], [1,100], [1,1000] words having the highest TF/IDF scores for that book. (3 features.)

With a trained model and learned weights for all features, we next featurize each sentence in a test book according to the same set of features described above and predict whether or not it will appear in the summary. Sentences are then ranked by probability and the top sentences are chosen to create a summary of 1,000 words. To create a summary, sentences are then ordered according to their position in the source document.

5 Evaluation

Document summarization has a standard (if imperfect) evaluation in the ROUGE score (Lin and Hovy, 2003), which, as an n-gram recall measure, stresses the ability of the candidate summary to recover the words in the reference. To evaluate the automatically generated summary, we calculate the ROUGE score between the generated summary and the held-out reference summary from Wikipedia for each book. We consider both ROUGE-1, which measures the overlap of unigrams, and ROUGE-2, which measures bigram overlap. For the case of a single reference translation, ROUGE-N is calculated as the following (where w ranges over all unigrams or bigrams in the reference summary, depending on N, and $c(\cdot)$ is the count of the n-gram in the text).

$$\frac{\sum_{w \in ref} \min(c(w_{ref}), c(w_{hyp}))}{\sum_{w \in ref} c(w_{ref})} \quad (9)$$

Figure 2 lists the results of a 10-fold test on the 439 available book/summary pairs. Both alignment models described above show a moderate improvement over the method of Jing et al. For comparison, we also present a baseline of simply choosing the first 1,000 words in the book as the summary.

Model	ROUGE-1	ROUGE-2
Block HMM	41.4	6.2
Word HMM	41.3	6.2
Jing 1999	40.7	6.0
First 1000	38.0	6.0

Table 2: ROUGE summarization scores.

How well does this method actually work in practice, however, at the task of generating summaries? Manually inspecting the generated summaries reveals that automatic summarization of books still has great room for improvement, for all alignment methods involved. Appendix A shows the sentences extracted as a summary for *Heart of Darkness*.

Independent of the quality of the generated summaries on held-out test data, one practical benefit of training binary log-linear models is that the resulting feature weights are *interpretable*, providing a data-driven glimpse into the qualities of a sentence that make it conducive to appearing in human-created summary. Table 3 lists the 25 strongest features predicting inclusion in the summary (rank-averaged over all ten training splits). The presence of a name in a sentence is highly predictive, as is its position at the beginning of a book (decile 0) or at the very end (decile 8 and 9). The strongest lexical features illustrate the importance of a character's persona, particularly in their relation with others (*father*, *son*,

etc.), as well as the natural importance of major life events (*death*). The importance of these features in the generated summary of *Heart of Darkness* is clear – nearly every sentence contains one name, and the most important plot point captured is indeed one such life event ("Mistah Kurtz – he dead.").

1. IS_NAME
2. DECILE_0
3. TF-IDF < 100
4. DECILE_8
5. mr.
6. TF-IDF < 10
7. father
8. love
9. son
10. brother
11. years
12. young
13. mother
14. family
15. DECILE_9
16. daughter
17. wife
18. man
19. boy
20. life
21. death
22. house
23. chapter
24. child
25. sir

Table 3: Strongest features predicting inclusion in a summary.

6 Conclusion

We present here two new methods optimized for aligning the full text of books with comparatively much shorter summaries, where the assumptions of the possibility of an exact word or phrase alignment may not always hold. While these methods perform competitively in a downstream evaluation, book summarization clearly remains a challenging task. Nevertheless, improved book/summary alignments hold intrinsic value in shedding light on what features of a work are deemed "summarizable" by human editors, and may potentially be exploited by tasks beyond summarization as well.

A Generated summary for *Heart of Darkness*

- " And this also , " said Marlow suddenly , " has been one of the dark places of the earth . " He was the only man of us who still " followed the sea . " The worst that could be said of him was that he did not represent his class .
- No one took the trouble to grunt even ; and presently he said , very slow – " I was thinking of very old times , when the Romans first came here , nineteen hundred years ago – the other day .... Light came out of this river since – you say Knights ?
- We looked on , waiting patiently – there was nothing else to do till the end of the flood ; but it was only after a long silence , when he said , in a hesitating voice , " I suppose you fellows remember I did once turn fresh - water sailor for a bit , " that we knew we were fated , before the ebb began to run , to hear about one of Marlow ' s inconclusive experiences .
- I know the wife of a very high personage in the Administration , and also a man who has lots of influence with , ' etc . She was determined to make no end of fuss to get me appointed skipper of a river steamboat , if such was my fancy .
- He shook hands , I fancy , murmured vaguely , was satisfied with my French .
- I found nothing else to do but to offer him one of my good Swede ' s
- Kurtz was ... I felt weary and irritable .
- Kurtz was the best agent he had , an exceptional man , of the greatest importance to the Company ; therefore I could understand his anxiety .
- I heard the name of Kurtz pronounced , then the words , ' take advantage of this unfortunate accident . ' One of the men was the manager .
- Kurtz , ' I continued , severely , ' is General Manager , you won ' t have the opportunity . ' " He blew the candle out suddenly , and we went outside .
- The approach to this Kurtz grubbing for ivory in the wretched bush was beset by as many dangers as though he had been an enchanted princess sleeping in a fabulous castle .
- In a moment he came up again with a jump , possessed himself of both my hands , shook them continuously , while he gabbled : ' Brother sailor ... honour ... pleasure ... delight ... introduce myself ... Russian ... son of an arch - priest ... Government of Tambov ... What ?
- Where ' s a sailor that does not smoke ? " " The pipe soothed him , and gradually I made out he had run

away from school , had gone to sea in a Russian ship ; ran away again ; served some time in English ships ; was now reconciled with the arch - priest .

- " He informed me , lowering his voice , that it was Kurtz who had ordered the attack to be made on the steamer .
- " We had carried Kurtz into the pilot - house : there was more air there .
- Suddenly the manager ' s boy put his insolent black head in the doorway , and said in a tone of scathing contempt : " ' Mistah Kurtz – he dead . ' " All the pilgrims rushed out to see .
- That is why I have remained loyal to Kurtz to the last , and even beyond , when a long time after I heard once more , not his own voice , but the echo of his magnificent eloquence thrown to me from a soul as translucently pure as a cliff of crystal .
- Kurtz ' s knowledge of unexplored regions must have been necessarily extensive and peculiar – owing to his great abilities and to the deplorable circumstances in which he had been placed : therefore – ' I assured him Mr .
- ' There are only private letters . ' He withdrew upon some threat of legal proceedings , and I saw him no more ; but another fellow , calling himself Kurtz ' s cousin , appeared two days later , and was anxious to hear all the details about his dear relative ' s last moments .
- Incidentally he gave me to understand that Kurtz had been essentially a great musician .
- I had no reason to doubt his statement ; and to this day I am unable to say what was Kurtz ' s profession , whether he ever had any – which was the greatest of his talents .
- This visitor informed me Kurtz ' s proper sphere ought to have been politics ' on the popular side . ' He had furry straight eyebrows , bristly hair cropped short , an eyeglass on a broad ribbon , and , becoming expansive , confessed his opinion that Kurtz really couldn ' t write a bit – ' but heavens ! how that man could talk .
- All that had been Kurtz ' s had passed out of my hands : his soul , his body , his station , his plans , his ivory , his career .
- And , by Jove ! the impression was so powerful that for me , too , he seemed to have died only yesterday – nay , this very minute .
- He had given me some reason to infer that it was his impatience of comparative poverty that drove him out there . " ' ... Who was not his friend who had heard him speak once ? ' she was saying .
- Would they have fallen , I wonder , if I had rendered Kurtz that justice which was his due ?

References

[Barzilay and Elhadad2003] Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Brown et al.1990] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16:79–85, June.

[Ceylan and Mihalcea2009] Hakan Ceylan and Rada Mihalcea. 2009. The decomposition of human-written book summaries. In *CICLing'09*, pages 582–593.

[Ceylan2011] Hakan Ceylan. 2011. *Investigating the Extractive Summarization of Literary Novels*. Ph.D. thesis, University of North Texas.

[Daumé and Marcu2005] Hal Daumé, III and Daniel Marcu. 2005. Induction of word and phrase alignments for automatic document summarization. *Comput. Linguist.*, 31(4):505–530, December.

[Dempster et al.1977] A. P. Dempster, M. N. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22.

[DeNero et al.2008] John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a bayesian translation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 314–323, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Jing and McKeown1999] Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 129–136, New York, NY, USA. ACM.

[Kazantseva and Szpakowicz2010] Anna Kazantseva and Stan Szpakowicz. 2010. Summarizing short stories. *Computational Linguistics*, 36(1):71–109.

[Lin and Hovy2003] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 71–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Marcu1999] Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 137–144, New York, NY, USA. ACM.

[Mihalcea and Ceylan2007] Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 380–389, Prague, Czech Republic, June. Association for Computational Linguistics.

[Och and Ney2003] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

[Osborne2002] Miles Osborne. 2002. Using maximum entropy for sentence extraction. In *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, AS '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Quirk et al.2004] Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain, July. Association for Computational Linguistics.

[Shen et al.2007] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 2862–2867, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Vogel et al.1996] Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Wei and Tanner1990] Greg C. G. Wei and Martin A. Tanner. 1990. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

[Yeh et al.2005] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng. 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Inf. Process. Manage.*, 41(1):75–95, January.