Automatic Summarization of Legal Decisions using Iterative Masking of Predictive Sentences

Linwu Zhong[*]
Language Technologies Institute
Carnegie Mellon University

Ziyi Zhong[*]
Language Technologies Institute
Carnegie Mellon University

Zinian Zhao[*]
Language Technologies Institute
Carnegie Mellon University

Siyuan Wang[*]
Language Technologies Institute
Carnegie Mellon University

Kevin D. Ashley
School of Law
University of Pittsburgh

Matthias Grabmair
Language Technologies Institute
Carnegie Mellon University

ABSTRACT

We report on a pilot experiment in automatic, extractive summarization of legal cases concerning Post-traumatic Stress Disorder from the US Board of Veterans' Appeals. We hypothesize that length-constrained extractive summaries benefit from choosing among sentences that are predictive for the case outcome. We develop a novel train-attribute-mask pipeline using a CNN classifier to iteratively select predictive sentences from the case, which measurably improves prediction accuracy on partially masked decisions. We then select a subset for the summary through type classification, maximum marginal relevance, and a summarization template. We use ROUGE metrics and a qualitative survey to evaluate generated summaries along with expert-extracted and expert-drafted summaries. We show that sentence predictiveness does not reliably cover all decision-relevant aspects of a case, illustrate that lexical overlap metrics are not well suited for evaluating legal summaries, and suggest that future work should focus on case-aspect coverage.

CCS CONCEPTS

• **Computing methodologies** → *Information extraction.*

KEYWORDS

legal case summarization, text classification

ACM Reference Format:
Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D. Ashley, and Matthias Grabmair. 2019. Automatic Summarization of Legal Decisions using Iterative Masking of Predictive Sentences. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3322640.3326728

1 INTRODUCTION

Each year, tens of thousands of veterans appeal their rejected application for disability benefits to the Board of Veterans' Appeals (BVA), which in turn renders decisions affirming, denying, or remanding the lower tier decision. Since 1991, the BVA has published an average of 55k decisions a year, most of which have been drafted by single judges and their support staff attorneys. The BVA has a long backlog and it is challenging, if not impossible, for the BVA, as well as appealing veterans and their legal counsel, to find patterns and examine coherence in this extraordinarily large collection.

We hope to help solve the sensemaking challenge with a system able to summarize decision-relevant aspects of a given BVA decision by selecting sentences that are predictive of the case's outcome. This is determined by a machine learning model trained on a sufficiently large corpus, thereby bypassing the need to manually construct a factor model of the domain and train/develop factor-specific language extractors/classifiers. The thematic focus, structural homogeneity, and size of the dataset support research on automatically summarizing legal cases using sentence vectorization and machine learning techniques, including neural networks.

Although the BVA is not a court in the judicial branch, its decisions have the character of legal judgments and follow the same structural patterns of justifying an order in a case by applying the governing law to the available evidence. BVA decisions employ the same case- and rule-based reasoning and citation patterns. In spite of the high caseload, its opinions are of considerable length and contain up to hundreds of sentences serving different purposes (i.e. evidence, findings, etc) but with a high level of redundancy. We seek to design a system that can automatically summarize legal case documents with key information by extracting sentences. To support legal practitioners, the summary should be coherent, representative, and provide different levels of details.

In this paper, we use a corpus of around 35,000 BVA cases concerning the single issue of disability compensation for service-connected Post-traumatic Stress Disorder (PTSD). We first employ a novel train-attribute-mask pipeline to iteratively select sentences from the case that are predictive for the outcome. Then we use a sentence type classifier and Maximum Marginal Relevance [2] to select summarization sentences from those predictive sentences, and finally fit them into a template.

A neural classification model can derive the contribution of each sentence to the case outcome prediction. Sentences with highest attribution scores can be incorporated as candidates. We hypothesize that the dataset is sufficiently large that fact patterns recur in

[*] The first four authors contributed equally to this research.

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6754-7/19/06…$15.00
https://doi.org/10.1145/3322640.3326728

multiple cases, and that the system can learn which fact patterns are associated with what case outcomes. Ideally, the sentences a trained model identified as predictive would cover all decision-relevant aspects of a case and hence be assembled into a summary.

As will be shown, while our predictive model achieves high accuracy during the sentence selection process, a comparison of the generated summaries to expert-created ones suggests that not all aspects deemed relevant by humans are captured by the system. We explore the summary data obtained using lexical overlap metrics and engage in comprehensive qualitative error analysis. In doing so, we make the following contributions to the state of the art in automatic summarization of legal documents:

- an iterative train-attribute-mask pipeline employing a convolutional neural network classifier to gradually extract predictive sentences from the opinion, which, as we demonstrate, is preferable over extracting the top k sentences from a single train-predict-attribute pass,
- experimental results of an extractive auto-summarization pipeline combining masking-based extraction, sentence classification, and maximum-information-based selection of sentences from decisions in both quantitative, and qualitative, comparison to expert-extracted, and expert-drafted, summaries of BVA decisions,
- a dataset[1] of (1) 92 single-annotated extractive 6-10 sentence summaries, created by 4 annotators using 6 types, of single-issue BVA cases focusing on PTSD, (2) a test set consisting of an additional 20 such cases quadruple-annotated for agreement evaluation, and (3) two summaries manually drafted by law students for each test set case.

2 RELATIONSHIP TO PRIOR WORK

2.1 General Summarization Techniques

Automatic text summarization techniques generally form two categories: abstractive [6, 15] and extractive. Our work belongs to the latter; document summaries are generated by computing the relative importance of sentences in the given document and selecting a subset. [4] use an intra-sentence similarity matrix as an adjacency matrix of graph representations of sentences and compute sentence importance based on the concept of eigenvector centrality in the graph. [7] discussed using Latent Semantic Analysis on a term-by-sentence matrix to select semantically important sentences, achieving good extractive summary coverage. The interaction between statistical and semantic features in extractive summarization was discussed in [21], which shows that semantic features like textual entailment benefit text redundancy detection, while statistical features like TF and IDF (i.e. term frequency and inverse document frequency) are able to select the most representative sentences in non-redundant text.

2.2 Legal Text Summarization

Prior work has applied extractive text summarization to legal texts. [10] experimented with a wide range of features and machine learning techniques to predict the rhetorical status of sentences and select the most summary-worthy sentences from judgments of the UK House of Lords. [5] built table-type summaries using documents' architecture and thematic structure to improve coherence and readability. By contrast, our work uses machine learning to first select which sentences in the decision are predictive for the outcome. We then partition suitable sentences using a type classifier, and select a set of summary sentences using maximum marginal relevance.

Some have applied graph based methods for legal text summarization. In [11] the authors treat sentences as nodes and represent documents as disconnected and directed graphs of sentences. Edges are computed by statistical methods, and each connected sub-graph is considered a cluster of the same topic in the document. The method in [17] constructs a graph representation of a legal document and uses a voting algorithm based on repetition of legal phrases as a sentence similarity measure.

2.3 Neural Text Classification

Neural networks can capture and learn semantic features from large amounts of text. We used neural network models with the case outcome as a supervision signal to extract sentences that contribute more to the prediction. First we needed to choose a neural network architecture. [1] applied a Hierarchical Attention Network (HAN) to predict the outcome of BVA decisions from their text and achieved decent results for long documents. A hierarchical attention network [23] is a multi-level encoder structure: from a word-level encoder to paragraph-level and lastly document-level, embeddings are generated hierarchically and input to a fully connected layer whose output is the classification result. A recurrent neural architecture (e.g. RNN, LSTM) is used across the individual levels of the HAN to capture sequential information and obtain the attention score of words and sentences. Convolution-based neural models (CNNs) have also been successful in various text classification contexts. [12] showed that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results over multiple benchmarks. By contrast to RNN and HAN models, CNN incorporates local information by using different convolution kernel sizes (i.e. the width of the 'filter' that passes over the text), and is faster to train because it does not use recurrent units.

Instead of learning sentence embeddings during training, as in HAN, we used the pre-trained Universal Sentence Encoder [3], a state-of-the-art sentence embedding model. It is trained using a variety of data sources and tasks. The input is variable length English text and the output is a 512 dimensional vector. The model is trained with the Transformer [20] encoder. For generality, it employs multi-task learning; a single encoding model feeds multiple downstream tasks, including a SkipThought like task [13], a conversational input-response task, and classification tasks.

2.4 Attribution for Deep Networks

[19] introduced a method to attribute the prediction of a deep network to its input features. The authors introduced the idea of "Integrated Gradients". They are derived by computing the linear integral from baseline input features to real input features in getting the final output value. The resulting single scalar represents the gradients, and essentially attributes the prediction to input features.

[1] Our experimental code is available at https://github.com/luimagroup/bva-summarization
We are planning to make as much of the experimental data available as possible.

Integrated Gradients are independent of the specific neural architecture and provide a measure of relevance for each sentence by quantifying its impact on the overall case outcome prediction.

2.5 Sentence Classification

Previous work in legal text summarization has explored how to use rhetorical roles of sentences in targeted content extraction by means of classification. [8, 9, 24] cluster sentences according to their rhetorical roles including FACT, BACKGROUND, and further sub-categories. We also employ a module that classifies sentences according to their 'function' in the legal text. Because the BVA corpus is highly regular, we only employ machine learning to classify 'Reasoning' and 'Evidential Support' sentences from other types. We rely on pattern-based extraction for other case information (i.e. 'Issue', 'Outcome', 'Procedural History' and 'Service History').

3 SUMMARIZATION TASK DESCRIPTION

Our goal is to generate summaries of single-issue BVA cases that are between 6-10 sentences long and adhere to the following structure:

- one sentence that identifies the issue on appeal,
- one sentence that summarizes where the appeal is from (e.g., a particular department, regional office, and city)
- one sentence that reports when/where the veteran served,
- one sentence that states the decision of the Board,
- depending on the length of the case, between two and six additional sentences that best summarize the Board's reasons for the decision and the evidence considered.

The standard for the summaries is that they should contain enough information for readers to be able to make an informed decision about whether to read the full decision given their interest in particular issues. Since we impose a restriction on the maximum number of sentences, our auto-summarization task is one of (1) extracting a pool of relevant sentences and (2) selecting a limited subset from this pool that maximizes the captured information.

4 DATA

We have obtained 972,522 BVA decisions[2] for appeals of rejected disability claims by US veterans from 1992 to 2017, around 100,000 of which are single issue cases. Around 35,000 of those deal with service-connected post-traumatic stress disorder (PTSD)[3] with a distribution of granted, denied and remanded cases of about 1:2:3.

We randomly sampled a dataset of single-issue PTSD decisions for this experiment. It comprised 112 cases where a veteran appealed a rejected initial claim for disability compensation to the BVA. Document lengths ranged from 564 to 6285 tokens with a mean of 1966 and a median of 1645 tokens. We created two kinds of case summaries: extractive and manually drafted.

Extractive Summaries: For the extractive summaries, four annotators with legal expertise (including the fifth and sixth author) annotated a stratified set of 92 summaries (each annotated 7 denied, 12 remanded, 4 granted cases) as training data, from which 24 cases were sample stratified as validation data (2 denied, 1 granted, 3 remanded from each annotator). The same group also annotated a randomly sampled test set of 20 cases (6 denied, 10 remanded, 4 granted) which doubled as the reliability corpus.[4]

We developed an annotation type system analogous to the above task description comprising the surface types 'Issue', 'Procedural History', 'Service History', 'Outcome' (each one sentence per decision), and two semantically complex types, 'Reasoning' and 'Evidential Support'. Annotators were instructed to annotate one sentence for each surface type.[5] A manual survey of the data had revealed that the decisions were sufficiently structured that the relevant information could usually be found in a single sentence at a predictable location. Then, starting from the outcome, the experts were asked to annotate between one and three 'Reasoning' sentences. These should state how the outcome is warranted (or not) by the available evidence in the case. These reasoning sentences connect the outcome to the facts, typically by assessing available or missing evidence, or persuasiveness/quality of evidence.[6] For example, if the appeal for benefits for service-connected PTSD was denied, then the 'Reasoning' sentences could state that the veteran did not succeed in proving that he has a current diagnosis of PTSD. Finally, the experts annotated 1-3 'Evidential Support' sentences that provide more information about the facts/evidence in the case. These 'Evidential Support' sentences add information about the factual basis on which the already annotated 'Reasoning' justifies the outcome. Facts that were somehow relevant but not connected to the reasoning were not annotated.

As a guiding principle, when in doubt about which sentences to include in the restricted sentence set, the annotators should strive to capture as much relevant information as possible. This type system is a compromise between permitting the annotators flexibility to capture the most informative sentences within length constraints and a fully structured premise-conclusion type system as used in argument mining [18].

Summary Agreement/Reliability: We did not calculate sentence selection agreement across extracted summaries because the high level of semantic redundancy in the documents would make the resulting statistic less informative. Two annotators picking different sentences with largely similar content should not count as disagreement. Instead we conducted ROUGE-score based lexical overlap analyses across annotators in section 6.3.

Drafted Summaries: Two first year law students (non-authors) each wrote summaries of the 20 test set cases as per instructions like the task description above. They were asked to stay within 6-10

[2] See the BVA website at https://www.bva.va.gov and searchrable decisions at https://www.index.va.gov/search/va/bva.jsp

[3] i.e. had a single issue sentence matching a regular expression targeted to detect compensation claims for service-connected PTSD

[4] Disclosure of sampling method: The team initially annotated a randomly sampled unstratified training set of 80 cases. At a later point we stratified the set by annotating additional randomly sampled cases, leading to the set size of 92, and moved already annotated surplus cases into the reliability set, which was then expanded and stratified with additional random cases from the overall collection.

[5] Disclaimer: For a small number of cases the sentence limit for individual types was sometimes exceeded. This typically occurred where the annotator believed that annotating only the permitted number sentences of a given type would risk conveying incorrect information (e.g. if the procedural history was more complex and involved transfers of jurisdictions). We did not correct this after data collection as those instances were rare and the information should not be discarded but rather kept to revise the annotation protocol. We believe the noise thus introduced to be acceptable.

[6] Our 'Reasoning' type corresponds closely to the 'Evidence Based Finding' in prior work on BVA decisions [22]. For this summarization task, our more abstractly framed type is slightly more intuitive for remanded cases where no finding on the merits is made but reasons are given for why the case is sent back to the Regional Officer.

sentences and not to exceed the number of words as determined by the longest extractive summary of each respective case.

Sentence Classification Data: Additionally, we have 26 BVA cases annotated at the sentence level with a type system comprises 'Citation', 'ConclusionOfLaw', 'Evidence', 'EvidenceBasedFinding', 'EvidenceBasedReasoning', 'Header', 'LegalPolicy', 'LegalRule', 'PolicyBasedReasoning', and 'Procedure'. This data stems in part from earlier, unpublished work by this group, and in part from our collaborators at Hofstra Law School's LLT Lab.

5 SYSTEM DESIGN

5.1 Extracting Predictive Sentences

In the first stage of our auto-summarization system (see Fig. 1) we extract each case's outcome (application for benefits granted, denied, or remanded) using regular expressions. For each decision, we first apply a legal text sentence segmenter [16] and filter out sentences that may statistically correlate with the case outcome but are not useful for inclusion in a summary (e.g. citations). We then use our train-attribute-mask pipeline to select predictive sentences from the text. The pipeline iteratively uses a CNN text classification model trained on the dataset using the case outcome as supervision and predicts whether the cases are granted, denied, or remanded. After the model converges, for each case in the training set, the trained model attributes the outcome prediction using the integrated gradient method. The sentence with the highest attribution score (i.e. the most predictive) is selected, added into our collection of relevant sentences for that case, and masked out in the decision text. The cycle then repeats with the model training on the set of training cases, each of which now has one sentence less. The procedure stops after 60 iterations, at which the prediction accuracy of the trained model on a validation set has dropped to around 0.71.[7]

5.2 Surveying Predictive Sentences

After briefly surveying unfiltered extracted predictive sentences for patterns for each class (granted, rejected, denied), we found that not all sentences are immediately recognizable as predictive for the outcome and/or suitable for summaries. Such sentences include bare citations to cases and statutes, apparently neutral statements about facts and procedure, sections headings, and seemingly random text fragments. We hypothesize that such statements can be predictive if they correlate with a higher probability of a certain case outcome than the base grant/deny/remand rate (e.g. certain precedents being cited more frequently for certain outcomes or BVA staff attorneys reusing identical text across multiple cases). As described in Sec. 5.1, our model remedies this by removing unsuitable sentences (citations, headers, short text artifacts, etc.) before applying the iterative masking procedure.

We also conducted a similar, brief survey of the sentences that had been extracted after filtering. Sentences that are readily apparent as predictive include statements of the case outcome, evidence-based findings, and evidence statements whose phrasing suggest a certain outcome. Extracted sentences whose predictive role is unclear include neutral statements of facts and procedure, legal rules and policies, and small numbers of citations and surface artifacts that the filters did not catch. Some of the extracted sentences did not seem predictive yet contained information useful for the summary (e.g. factual statements about traumatic events suffered by the veteran). A full survey of extracted sentences for all case outcomes, and analysis of why they were extracted as predictive, exceeds the scope of this experiment and may be pursued in future work. In particular, we would like to examine whether sentences extracted from cases that are mispredicted at certain masking iterations are less suitable for summaries. One possible improvement to the model could then be to stop extracting sentences from case texts when the case outcome can no longer be correctly predicted, or leverage information about which sentences are influential in the model's misprediction.

5.3 CNN-based Text Classification

We chose to encode case documents via the Universal Sentence Encoder (USE) [3]. The USE is pre-trained on general text corpora and generates sentence-level embeddings. Inspired by the work of [12], we use a sentence-n-gram CNN model to perform the case outcome prediction task. The USE encodes each ith sentence in a document into a k-dimensional sentence embedding $s_i \in R^k$. The document of length n is hence represented as a $n \times k$ embedding matrix $s_{1:n}$, where $_{i:j}$ corresponds to the embedding sub-matrix of sentences s_i, $_{i+1}$, $_{i+2}$, ..., s_j. This is similar to the idea of token n-grams, but with sentences being the atomic unit. The convolutional layers then apply filters to windows of h sentences to capture composite patterns as higher order features. Eventually, the outputs of these filters are aggregated as filter-specific feature maps. We employ window sizes h of 2, 3 and 4, and train 256 filters for each window size. After performing the convolution operations of different window size options in parallel, we apply a global-max pooling operation over the feature maps of each option and then do a concatenation. Eventually, the concatenated vector runs through a dense layer to predict the case outcome.

Several considerations led to our decision to use the n-gram CNN as the text classification model, instead of a recurrent HAN. CNNs outperform recurrent models in terms of training time, which is essential for our iterative train-attribute-mask pipeline. They can also be easily combined with our Integrated Gradients attribution method because we are able to directly attribute back to each sentence and determine the most predictive one after each round. With a recurrent HAN, we would need to attribute back to each word and aggregate word attributions sentence-wise.

5.4 Showing that Iterative Masking Works

Our iterative masking procedure is based on the assumption that the CNN model will initially place higher predictive weight on the obvious statements of the case outcome in the text, but learn to pick up more subtle expressions as the former are increasingly obfuscated at training time. We validate this assumption empirically by running the iterative masking procedure without filtering sentences unsuitable for summaries beforehand. First, we compare the per-round prediction accuracy of the first round trained model with the iteratively retrained one. As shown in Fig. 2, the per-round retrained model can maintain an over 0.8 F1 outcome prediction

[7] For decisions in the collection that were 60 sentences or shorter, this results in all sentences being considered for the summary.

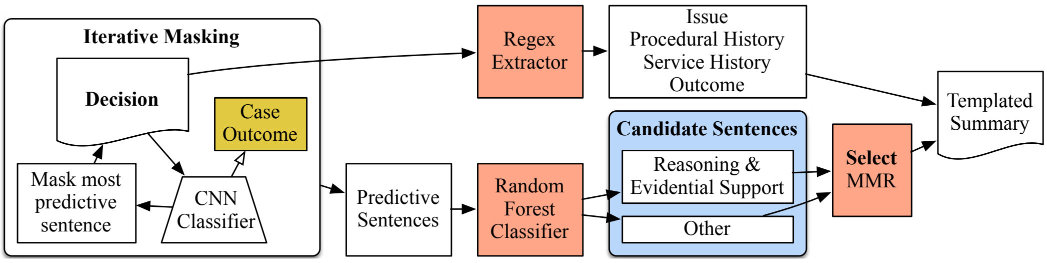


Figure 1: Overview of the Summarization System

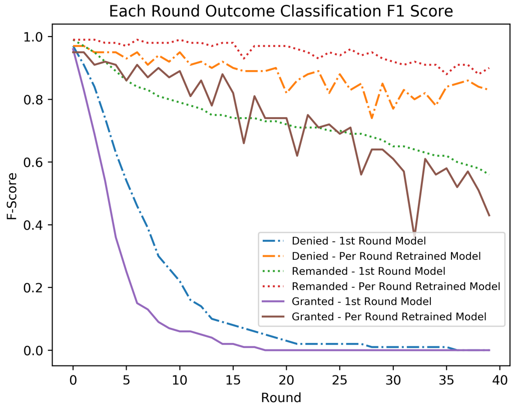


Figure 2: Per-round Outcome Prediction F1 Score on Masked Case Text Using First Round Model vs Per-Round Re-trained Models (164 Granted, 309 Denied, 483 Remanded)

score until 20 sentences (or even more) have been masked, varying by the case outcome. However, the performance of the first round trained model drops to below random guessing very quickly. This deterioration confirms that the attribute-and-mask procedure can identify and mask out predictive sentences effectively. Moreover, it shows that the per-round retrained model is indeed able to learn from the remaining sentences and pick up signal that the first-round-only trained model misses.

Second, as further evidence of this, Fig. 3 contains a histogram of the cases in our training set showing the size of the set of sentences shared between the top 40 attributed sentences from the first round model and the masked sentences from all 40 rounds using the per-round retrained model. It illustrates that, for almost all cases, iteratively masking one sentence and re-training at each iteration predicts a different set of sentences as predictive compared to selecting multiple sentences (in this case 40) in the first round.

5.5 Sentence Type Classification

Once extracted, predictive sentences are then classified as two types: 'Reasoning/EvidentialSupport' (i.e. the union of the two types from the annotation scheme) or 'Other'. To help determine if a sentence

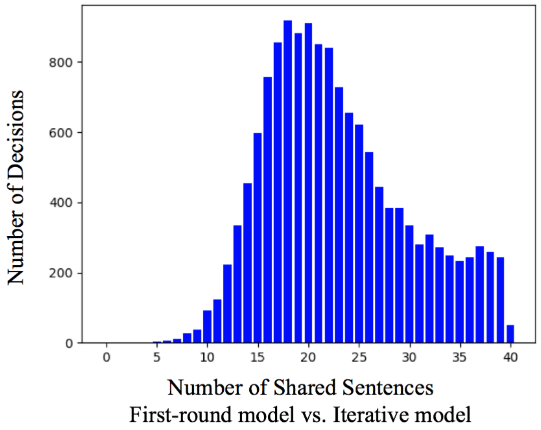


Figure 3: Histogram of Overlap between Top 40 Attributed Sentences from First-Round and Per-round retrained model

in a case document is a 'Reasoning/EvidentialSupport' sentence, we generated a small number of feature extraction rules based on a very small labeled development dataset.[8] We chose a random forest with 10 estimators/trees as our model. The heuristic features are concatenated with the embeddings obtained from the USE and fed into the model as input features. We constructed a classification dataset by composing positive and negative sentence instances from our data. First, we downsampled sentences from the pre-existing dataset of 26 sentence-annotated BVA decisions that were not annotated as 'EvidenceBasedReasoning' or 'EvidenceBasedFinding' to be used as negative examples. Then, for positive examples, we used all 'Reasoning' and 'Evidential Support' sentences marked up in the 72 summarization training set cases for training, and the same partition of the 24 validation set cases for testing. Our classification dataset eventually comprised 954 training sentences (579 positive, 375 negative) and 341 testing sentences (216 positive samples, 125 negative). Our random forest classifier reached .85 precision, .77 recall, and .81 F for distinguishing the joint 'Reasoning/EvidentialSupport' type from the 'Other' types.

[8] Features: Number of words in sentence; percentage of uppercase letters in sentence; fractions of noun, verb and preposition POS tagged tokens; fraction of years and numbers among all tokens; number of periods in sentence, binary existence of some clue words (e.g. "according", "likely", "example")

5.6 Summary Sentence Selection

After selecting predictive sentences and classifying them into 'Reasoning/EvidentialSupport' and 'Other', we apply maximum marginal relevance with TFIDF sentence representation in the group of 'Reasoning/EvidentialSupport' sentences to select sentences for the summary. If not enough sentences of this type are available, the 'Other' category is considered as well. 'Issue', 'Procedural History', 'Service History', and 'Outcome' sentences are extracted by regular expression from the case document. Using a variable-length summary template, the final summary consists of the following:

(1) A sentence stating the procedural history, e.g. "This is an appeal from the Department of Veterans Affairs Regional Office in Seattle, Washington."
(2) A sentence describing the issue of the case, e.g. "The issue is entitlement to service connection for posttraumatic stress disorder (PTSD)."
(3) A sentence describing the service history of the veteran, e.g. "The veteran had active military service from November 1967 to December 1970."
(4) Variable number of 'Reasoning/EvidentialSupport') sentences selected by classification and MMR from predictive sentences restricted to the average number of 'Reasoning' and 'Evidential Support' sentences in the human-annotated summaries, rounded up.
(5) A sentence asserting the conclusion, e.g. "Service connection for PTSD is granted."

5.7 Maximal Marginal Relevance

Maximum Marginal Relevance (MMR) [2] is a widely used criterion for diversifying sentence selection in summarization. As shown in the formula, MMR ranks sentences according to their relevance scores, as well as reduces redundancy between selected sentences:

$$MMR(S_i) = \lambda \times Sim(S_i, Case) - (1 - \lambda) \times Sim(S_i, Summary)$$

Given a collection of 'Reasoning/EvidentialSupport' sentences, a sentence S_i with the maximum marginal relevance score will be iteratively removed and added to the summary. *Case* refers to the entire legal document, while *Summary* refers to the selected sentences that have been included in the summary. We use cosine similarity to measure how similar a document (or sentence) is, represented by a single TFIDF vector.[9] We tune the parameter λ to find the best ratio of sentence relevance to novelty (sec. 6.2).

6 RESULTS

6.1 Automatic Summarization

Ideally, evaluating a text summarization system involves assessing summary fluency and adequacy. ROUGE is widely used [14] as an evaluation metric for automatic summarization tasks. We use measures based on the number of overlapping unigrams (ROUGE-1) and bigrams (ROUGE-2) between the computer-generated summary to be evaluated and the ideal summaries created by humans.

To assess our iterative masking procedure in Fig. 1, we compared system-generated and human-generated summaries of the 24 validation documents. We also generated baseline summaries without the train-attribution-mask pipeline by classifying all sentences of the document, instead of only predictive sentences, applying MMR on each group, and then fitting them into the summary template. Table 1 shows the average ROUGE score of the generated summaries side by side with and without the iterative-masking based selection. For this condition we set the λ MMR parameter to 0.5.

For this validation experiment, the ROUGE scores are calculated only on the 'Reasoning/EvidentialSupport' sentences instead of the whole summary. This is because we can achieve perfect overlap for the 'Issue', 'Procedural History', 'Service History', and 'Outcome' sentences since they would be extracted by the same regular expressions in either summary. The results show that the train-attribute-mask pipeline improves at least 2 points in all ROUGE metrics. This could be considered evidence for our hypothesis that the case outcome predictiveness of a sentence is a proxy for its relevance in an extractive summary.

As a counterpoint, we found that 32% of 'Reasoning/Evidential Support' sentences from the human labeled summaries in the training data are not selected by the iterative extraction process, which is a fair amount. Because of the high redundancy of legal documents, sometimes sentences that are equally appropriate for inclusion in the summary are arbitrarily chosen by the annotator. Hence, in terms of legal text summarization, a ROUGE score may not perfectly reflect the actual quality of the summary. This is further supported by the full quantitative result comparison below in sec. 6.3.

6.2 Redundancy vs. Information Gain

Automatic Scoring: We also examined the impact of the MMR λ parameter on the quality of generated summaries. λ refers to the weight MMR puts on relevance versus diversity when ranking sentences. If λ is 1.0, MMR picks sentences based on the similarity (relevance) between each sentence and the whole document; if λ is 0.0, MMR picks sentences that maximize the diversity among selected sentences. Therefore, we generated summaries on the 68 training documents using the λ ranges from 0.1 to 0.9. We found that, surprisingly, forcing MMR to choose more diversified sentences (smaller λ value) leads to lower ROUGE scores for the generated summaries. By examining the generated summaries for each case under different λ values, we observed that the selected 'Reasoning/EvidentialSupport' sentences are more similar to each other when λ increases. While this leads to more "redundancy", it also facilitates the selection of sentences that address similar aspects of the case. Smaller values of λ may result in the retrieval of sentences containing rare words from the case document (since they are diverse) as determined by the TFIDF representation. In fact, we observe that including diverse sentences can sometimes deteriorate the coherence and logical structure of a generated summary as they distract the summary from the main narrative. In other words, a good summary can add relevant information even by using common words that are closer together in TFIDF space.

Expert Relative Ranking: To further investigate this, the fifth author (who is a law professor) manually compared the 20 generated test set summaries for $\lambda = .2, .5, .8$ side by side with the two expert

[9] Our group has observed in prior, unpublished work on case summarization that TFIDF vectorization for MMR can produce summaries that humans score at least as highly as ones produced using aggregated, specially trained word embeddings. As a qualitative evaluation of summaries generated by multiple vectorization approaches would be prohibitively time intensive, we hence chose TFIDF as the single sentence representation for the summary composition experiments presented here.

	w/ train-attribute-mask pipeline	w/o train-attribute-mask pipeline
ROUGE-1	**0.269** (±**0.171**)	0.233 (±0.158)
ROUGE-2	**0.102** (±**0.178**)	0.082 (±0.150)

Table 1: ROUGE Score of Summarization

drafted ones. We also added a control summary whose sentences were randomly selected from the filtered predictive sentences without performing sentence type classification and MMR, but included the 'Issue', 'Procedural History', 'Service History', and 'Outcome' sentences. He then determined a relative ranking of the three generated summaries and the random selection summary, from best (1) to worst (4). The best scoring condition was $\lambda = .8$ with an average rank of 1.95, followed by $\lambda = .5$ with average rank of 2.53, $\lambda = .2$ with 2.74, and random as worst with an average rank of 2.79. These results provide further support that redundancy in the sentence selection helps narrative cohesion. For the quantitative and qualitative evaluation in sec. 6.3 and 6.4, we hence generate summaries with $\lambda = .8$. In light of these results on the effects of summary diversification we would like to explore different vectorizations and diversification strategies for automatic summarization in future work by using, for example, word embeddings or language models.

6.3 Quantitative Evaluation

We conducted a quantitative evaluation of generated summaries using ROUGE-1 (unigram overlap) and ROUGE-2 (bigram overlap). Table 2 shows the results. In each cell, the row-specific summary is considered the experimental one (the hypothesis) and compared to the column-specific summary as the reference. These scores are generally higher than in the experiment above in sec. 6.1 as they are based on the full summary including the regex-extracted sentences. This is because we consider the human-drafted summaries to be the standard for scoring and are also interested in how close human-extracted summaries come to hand-drafted ones. From this table of scores, we can make a series of observations.

Comparing drafted summaries: An immediately visible result is that the two human drafted summaries only reach scores of .73/.72 ROUGE-1 and .55/.54 ROUGE-2 in a pairwise comparison. Assuming that both drafted summaries are of high quality, this suggests that the ability of ROUGE scores to serve as measures of correctness for this summarization task on comparatively long summaries may be limited. We assume this is due to ROUGE's reliance on lexical overlap and failure to capture deeper semantic similarity. Hence, we conducted an extended qualitative analysis of all summaries below in sec. 6.4 (except for the random predictive condition).

Extracted vs. Extracted: The extracted summaries by the four annotators score fairly high in both ROUGE-1 (ranging from .88 to .95) and ROUGE-2 (from .85 to .92) when compared to one another, which is not unexpected since they are choosing sentences from the same pool and are not engaging in any paraphrasing.

Extracted vs. Drafted: Three out of four extractive annotators have higher ROUGE scores in comparison to Drafter 1's summaries than Drafter 2's summaries have in comparison to Drafter 1. This suggests that Drafter 1 engaged in less paraphrasing than Drafter 2 and possibly transferred phrases unchanged while writing.

Generated vs. Random: MMR-selected predictive sentences and randomly selected predictive sentences produce nearly identical ROUGE scores when compared against drafted (.73 and .79 ROUGE-1, .52 and .62 in ROUGE-2) and annotated summaries (.63-.68 ROUGE-1 and .44-.52 ROUGE-2). In comparison to drafted summaries, they tend to score higher than manually extracted summaries for Drafter 2, yet lower for Drafter 1. This is further evidence that the two sets of manually drafted summaries differ significantly and (assuming that both are good summaries) further limits the use of automatic metrics for this task. Interestingly, generated summaries tend to score higher overall in comparison to drafted summaries than to extracted summaries, which is counterintuitive since they have also been composed by selecting sentences from the document.

6.4 Qualitative Error Analysis

To get a better sense of the summaries' quality, we undertook a manual qualitative analysis.

As an example comparison, Fig. 4 illustrates three summaries of case no. 16. A first year law student generated the first (Human), a human with legal training generated the second by extracting sentences from the case by hand (SEBH), and the machine generated the third (MG). All three mention the Board's request on remand for an additional medical opinion of whether the veteran's service-connected PTSD caused or exacerbated his hypertension.

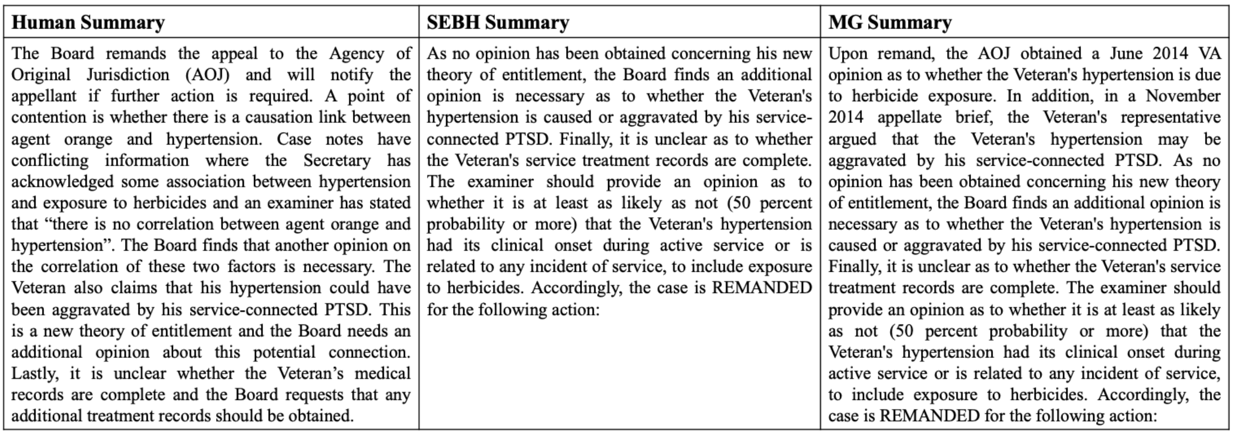
For each of 20 cases, the fifth author (a law professor) examined a set of seven summaries: the two drafted summaries, the four summaries extracted by hand, and a machine-generated summary from our system with an MMR $\lambda = .8$. The task was to compare the MG summary with the SEBH and Human summaries to identify any patterns of "errors," that is, important features of the Human or SEBH summaries that the MG summary lacked, or features of the MG summary that were extraneous in comparison to those in the Human or SEBH summaries. Good case summaries should, at least, capture the important legal issue(s) in a case and their resolutions. For each of the twenty cases, Figure 5 sets forth the important issue, its resolution, and the number of summaries of each type (Human, SEBH, MG) that adequately identify the issue and resolution.

As shown, the MG summaries adequately identified the main legal question in only half of the cases. The two Human summaries and four SEBH summaries of case no. 1, all captured that there were conflicting medical opinions as to whether the veteran's psychological condition did or did not meet the criteria of PTSD, and the fact that the Board resolved the reasonable doubt in favor of the veteran by applying the evidence-in-equipoise rule.

For an MG summary to be regarded as adequate, we required that it *expressly* describe the issue and its resolution. The MG summary in case no. 1 referred to the conflicting evidence being in equipoise, but it did not expressly mention the Board's resolution of the reasonable doubt in favor of the veteran (although a reader might infer that from the outcome reported in the summary).

Hypothesis\Reference	Drafter 1	Drafter 2	Annotator 1	Annotator 2	Annotator 3	Annotator 4
Drafter 1	1.0 / 1.0	0.72 / 0.54	0.75 / 0.59	0.69 / 0.52	0.74 / 0.58	0.74 / 0.56
Drafter 2	0.73 / 0.55	1.0 / 1.0	0.67 / 0.44	0.62 / 0.4	0.67 / 0.46	0.64 / 0.41
Annotator 1	0.8 / 0.62	0.7 / 0.47	1.0 / 1.0	0.89 / 0.85	0.94 / 0.92	0.95 / 0.92
Annotator 2	0.72 / 0.54	0.64 / 0.41	0.88 / 0.85	1.0 / 1.0	0.85 / 0.8	0.9 / 0.89
Annotator 3	0.78 / 0.61	0.7 / 0.47	0.93 / 0.91	0.85 / 0.81	1.0 / 1.0	0.91 / 0.88
Annotator 4	0.77 / 0.59	0.67 / 0.42	0.94 / 0.92	0.91 / 0.89	0.91 / 0.87	1.0 / 1.0
generated $\lambda = .8$	0.73 / 0.52	0.79 / 0.62	0.66 / 0.48	0.62 / 0.45	0.68 / 0.5	0.63 / 0.44
random predictive	0.73 / 0.52	0.79 / 0.62	0.67 / 0.5	0.63 / 0.46	0.68 / 0.52	0.64 / 0.45

Table 2: ROUGE-1/ROUGE-2 summary scores (with column text as reference) averaged across test cases



Human Summary	**SEBH Summary**	**MG Summary**
The Board remands the appeal to the Agency of Original Jurisdiction (AOJ) and will notify the appellant if further action is required. A point of contention is whether there is a causation link between agent orange and hypertension. Case notes have conflicting information where the Secretary has acknowledged some association between hypertension and exposure to herbicides and an examiner has stated that "there is no correlation between agent orange and hypertension". The Board finds that another opinion on the correlation of these two factors is necessary. The Veteran also claims that his hypertension could have been aggravated by his service-connected PTSD. This is a new theory of entitlement and the Board needs an additional opinion about this potential connection. Lastly, it is unclear whether the Veteran's medical records are complete and the Board requests that any additional treatment records should be obtained.	As no opinion has been obtained concerning his new theory of entitlement, the Board finds an additional opinion is necessary as to whether the Veteran's hypertension is caused or aggravated by his service-connected PTSD. Finally, it is unclear as to whether the Veteran's service treatment records are complete. The examiner should provide an opinion as to whether it is at least as likely as not (50 percent probability or more) that the Veteran's hypertension had its clinical onset during active service or is related to any incident of service, to include exposure to herbicides. Accordingly, the case is REMANDED for the following action:	Upon remand, the AOJ obtained a June 2014 VA opinion as to whether the Veteran's hypertension is due to herbicide exposure. In addition, in a November 2014 appellate brief, the Veteran's representative argued that the Veteran's hypertension may be aggravated by his service-connected PTSD. As no opinion has been obtained concerning his new theory of entitlement, the Board finds an additional opinion is necessary as to whether the Veteran's hypertension is caused or aggravated by his service-connected PTSD. Finally, it is unclear as to whether the Veteran's service treatment records are complete. The examiner should provide an opinion as to whether it is at least as likely as not (50 percent probability or more) that the Veteran's hypertension had its clinical onset during active service or is related to any incident of service, to include exposure to herbicides. Accordingly, the case is REMANDED for the following action:

Figure 4: Example summaries drafted by a law student, extracted by an expert, and machine-generated

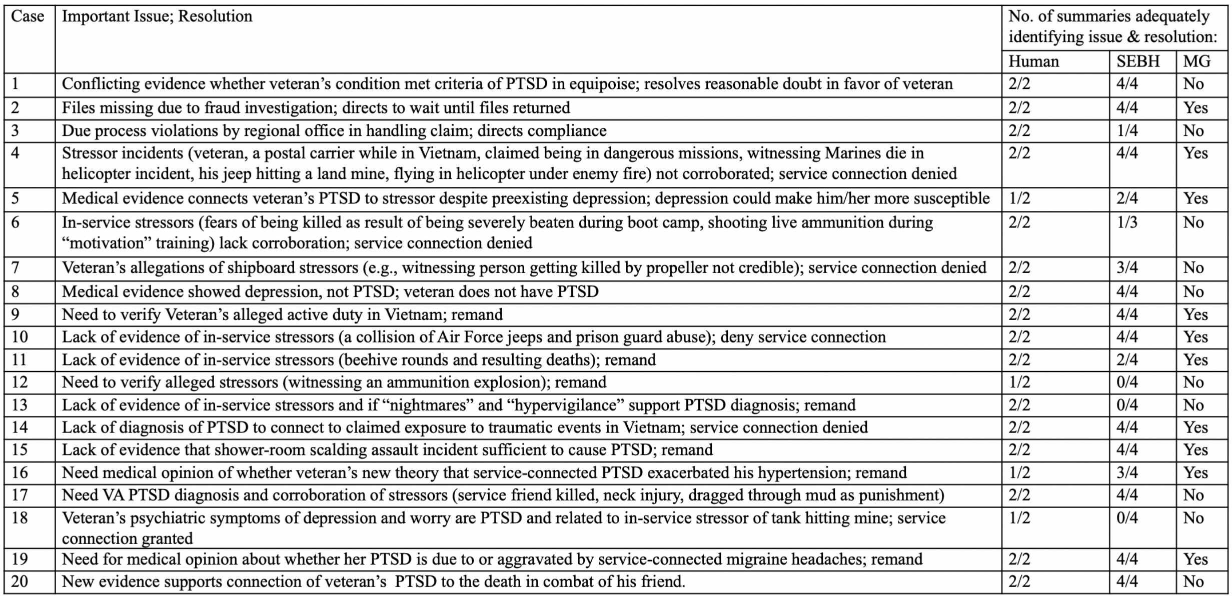
As noted, we hypothesized that a good summary should include the most predictive sentences. And yet, a good summary is a narrative; it needs to tell a story and provide some relevant details. For example, what the veteran experienced in the military that allegedly caused his or her condition may be a key indicator of its relevance in a new scenario involving the same kinds of causes-and-effects. Specific details, however, are not likely to be very predictive given their uniqueness, and thus our approach without more would be unlikely to select them unless they happen to fall within an otherwise predictive sentence. For instance, the SEBH summary #2 in case no. 1 reports that the veteran complained of recalling events that caused his psychological state such as "witnessing a fellow soldier who fell from a helicopter, and killing four Vietnamese with a grenade." The human sentence extractor rightly regarded those details as important. The MG summary (and three of the SEBH summaries) lacks that potentially useful specific information.

Such specific information may bear on the issue in subtle ways, some of which, ideally, an automated summarizer would catch. For example, they may relate to ways to reason about types of evidence or to evaluate witness statements:

Reasoning about types of evidence: In case no. 14, one Human and four SEBH summaries, but not the MG summary, mentioned the lack of Social Security treatment records, a common kind of evidence in these matters often missing from the record. In case no. 9, one Human, one SEBH *and* the MG summary all refer to buddy statements and Congressman's statements as a kind of evidence that is insufficient to establish war service.

Common sense experience in evaluating witness statements: In case no. 7, for instance, the veteran's testimony regarding a shipboard incident is noted as being inconsistent with his prior testimony, deck logs, and medical records. This inconsistency contributed to the Board's finding of the veteran's lack of credibility. These inconsistencies were mentioned in one Human and the SEBH summaries but not in the MG summary. In case no. 6 one of the Human summaries and one of the auto-generated summaries but not this particular MG summary, focused on the fact that the veteran was unable to remember the names of victims or other participants in the alleged stressor events, again an indicator of lack of credibility.

Some of this reasoning about witness statement specifics can be quite subtle. In case no. 4, for instance, both Human summaries included information that the veteran, a postal carrier while in Vietnam, had submitted written statements about being in dangerous missions, including witnessing Marines die in a helicopter incident, his jeep hitting a land mine, and riding in a helicopter coming under enemy fire. While the MG and the four SEBHs mentioned the need for corroboration of the veteran's statements, none described what the statements were about. The details of these statements of a *mail carrier* (as opposed to an infantryman or a pilot) underscored



Case	Important Issue; Resolution	No. of summaries adequately identifying issue & resolution:		
		Human	SEBH	MG
1	Conflicting evidence whether veteran’s condition met criteria of PTSD in equipoise; resolves reasonable doubt in favor of veteran	2/2	4/4	No
2	Files missing due to fraud investigation; directs to wait until files returned	2/2	4/4	Yes
3	Due process violations by regional office in handling claim; directs compliance	2/2	1/4	No
4	Stressor incidents (veteran, a postal carrier while in Vietnam, claimed being in dangerous missions, witnessing Marines die in helicopter incident, his jeep hitting a land mine, flying in helicopter under enemy fire) not corroborated; service connection denied	2/2	4/4	Yes
5	Medical evidence connects veteran’s PTSD to stressor despite preexisting depression; depression could make him/her more susceptible	1/2	2/4	Yes
6	In-service stressors (fears of being killed as result of being severely beaten during boot camp, shooting live ammunition during “motivation” training) lack corroboration; service connection denied	2/2	1/3	No
7	Veteran’s allegations of shipboard stressors (e.g., witnessing person getting killed by propeller not credible); service connection denied	2/2	3/4	No
8	Medical evidence showed depression, not PTSD; veteran does not have PTSD	2/2	4/4	No
9	Need to verify Veteran’s alleged active duty in Vietnam; remand	2/2	4/4	Yes
10	Lack of evidence of in-service stressors (a collision of Air Force jeeps and prison guard abuse); deny service connection	2/2	4/4	Yes
11	Lack of evidence of in-service stressors (beehive rounds and resulting deaths); remand	2/2	2/4	Yes
12	Need to verify alleged stressors (witnessing an ammunition explosion); remand	1/2	0/4	No
13	Lack of evidence of in-service stressors and if “nightmares” and “hypervigilance” support PTSD diagnosis; remand	2/2	0/4	No
14	Lack of diagnosis of PTSD to connect to claimed exposure to traumatic events in Vietnam; service connection denied	2/2	4/4	Yes
15	Lack of evidence that shower-room scalding assault incident sufficient to cause PTSD; remand	2/2	4/4	Yes
16	Need medical opinion of whether veteran’s new theory that service-connected PTSD exacerbated his hypertension; remand	1/2	3/4	Yes
17	Need VA PTSD diagnosis and corroboration of stressors (service friend killed, neck injury, dragged through mud as punishment)	2/2	4/4	No
18	Veteran’s psychiatric symptoms of depression and worry are PTSD and related to in-service stressor of tank hitting mine; service connection granted	1/2	0/4	No
19	Need for medical opinion about whether her PTSD is due to or aggravated by service-connected migraine headaches; remand	2/2	4/4	Yes
20	New evidence supports connection of veteran’s PTSD to the death in combat of his friend.	2/2	4/4	No

Figure 5: Overview of comparative summary error analysis

the need for corroboration. In case no. 10, a Human and one SEBH summary mentioned two alleged stressors, a collision of Air Force jeeps and prison guard abuse. The MG summary mentioned the latter in reporting insufficient corroboration but not the former. Presumably a collision of military jeeps is the type of event about which the military keeps records.

Novel theories and causation paths: In case no. 5, one Human, two SEBH, *and* the MG summary mentioned medical evidence that connects the veteran’s PTSD to a stressor despite his preexisting depression; instead of negating causation, his depression could make him *more* susceptible to such an incident thus supporting the causal connection. In case no. 16, one Human, four SEBH and, again, the MG summary mentioned the issue of the hypersensitivity theory (aka the “eggshell" theory), which the Human summary suggests, is a key issue to be resolved on remand.

Sometimes, the specific details are necessary to prevent these narratives about former service people, all suffering from psychological problems like PTSD, from becoming generic and anodyne. After all, the “T” in PTSD stands for human *trauma.*

Narratives about PTSD in military context: In case no. 11, one Human summary and the MG summary mentioned the “beehive rounds and resulting deaths" in emphasizing the need for corroboration and lack of VA efforts to corroborate. In case no. 12 one of the humans thought it worth mentioning the alleged stressor, witnessing an ammunition explosion. None of the other summaries mentioned it, including the MG summary. In case no. 13, two Human summaries mentioned “nightmares" and “hypervigilance", words the MG summary did not employ. In case no. 17, the MG summary did not mention the three specific stressors referred to in the Human summaries: the veteran’s service friend being killed, the veteran’s experiencing a neck injury in connection with military equipment, and the veteran’s being dragged through the mud as punishment for not protecting his weapon. None of the SEBH summaries mentioned these either. In case no. 18, the MG’s references to a stressor did not specifically mention the veteran’s tank hitting a mine.

Since the automated summary generator does not understand the sentences in depth, the fact that the MG summary occasionally contains desirable specific details is largely accidental. It is not likely that the summarizer will be able to detect details relevant to narratives about PTSD in military contexts; this would require too high a level of semantic understanding, a very knowledge-engineering-intensive domain mode or, possibly, too large a set of BVA case data for machine learning. On the other hand, if the program can identify sentences involving evidential reasoning, a role we do expect an ML program can learn to identify, it may have some basis for including in automated summaries details in sentences that involve reasoning about types of evidence or evaluating credibility of witness statements.

Interestingly, certain concepts leap to the attention of human summarizers that an automated summary may miss. In case no. 2, the Board determined that it would not be appropriate to proceed further until a pending fraud investigation’s claims files had been returned. While all of the summaries focused on the remand in order to obtain the information from the claims files, both Human summaries and two of four SEBH summaries explicitly mentioned the “pending *fraud* investigation" while the MG summary did not. The concept of “fraud" seemed salient to the humans even though the key fact was that the files temporarily were missing.

Some of the MG summaries contained material that seemed extraneous to good summaries, such as: (1) medical details of the PTSD diagnostic criteria [case no. 5], (2) issues unrelated to PTSD to be dealt with in separate opinions [case no. 6], (3) boilerplate assurances of the VA's complying with obligations to notify or assist the veteran [case no. 8], (4) restatements of the three PTSD service connection requirements [case no. 10], (5) detailed lists of medical examinations, dates, and diagnoses [case no. 15].

7 DISCUSSION

Our experiments evaluated the use of a sentence's outcome predictiveness, as determined by an iterative masking pipeline, as a measure of relevance and suitability for inclusion in an extractive summary. The motivation was in part to avoid the need to (1) construct and maintain a domain model of the relevant law and its application to cases (in, e.g., an ontology) and to (2) develop or train classifiers reliably connecting such a model to case texts. Our research goal was to discover if outcome-predictiveness in our dataset can serve as a proxy for such domain-model-like information. After examining the results, the answer to these questions is a qualified 'no', but some insight has been gained. To begin, we have produced evidence that automatic evaluation metrics based on lexical overlap, such as ROUGE, are of limited use for evaluating summaries of legal opinions as even summaries drafted by humans have low overlap scores. We have engaged in a qualitative comparison of the generated summaries with extracted, as well as drafted, ones and observed that the quality of summaries varies very much in terms of what we may label 'aspect-recall'. Cases have a set of decision-relevant aspects which seems to only partially overlap with sentences that are predictive for the outcome.

At the same time, we cannot treat or evaluate aspect extraction as a straightforward span classification problem since aspects are addressed in multiple locations across the decision. One would need an annotation scheme that treats distinguishable reasons for the outcome as first class objects and then links them to text spans where they are addressed by the judge/court. This moves automatic summarization yet closer to data structures used in argument mining. Since individual argument objects may be verbalized in more than one location, catching even one in an extractive summary can inform an aspect-focused precision/recall statistic that could guide supervised summarization models.

8 CONCLUSIONS & FUTURE WORK

Our experiments have shown that extracting sentences that are predictive for the case outcome by iteratively masking them and retraining the model at every round maintains higher prediction accuracy than models that are trained only once, presumably leading to a better extraction of predictive sentences. While this procedure improves lexical overlap metrics of generated summaries over non-predictive sentence selection, our qualitative analysis shows that the model misses some aspects of the cases that humans deem relevant. In future work we would continue to explore how large collections of cases enable a data-driven inference about which parts of the text are relevant for the case outcome to reduce the need for domain-specific model building. We believe this must involve both different (unsupervised) text representations (word embeddings, language models, etc.) and a dataset design that better supports the use of quantitative evaluation metrics by adapting ideas from, and moving our work closer to, argument mining.

9 ACKNOWLEDGMENTS

The authors thank University of Pittsburgh Intelligent Systems Program students Huihui Xu and Jaromir Savelka for producing extractive summary data, Law Students Alexandra Ashley and Stacy Yerington for drafting the hand-written summaries, and Vern Walker's group at the Law, Logic and Technology Lab at Hofstra University Law School for sharing their BVA sentence annotation dataset.

REFERENCES

[1] L. K. Branting, A. Yeh, B. Weiss, E. Merkhofer, and B. Brown. 2015. Inducing predictive models for decision support in administrative adjudication. In *AI Approaches to the Complexity of Legal Systems.* 465–477.

[2] J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proc. SIGIR 1998*, 335–336.

[3] D. Cer, Y. Yang, S.-Y.i Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil. 2018. Universal Sentence Encoder for English. In *Proc. EMNLP 2018*. ACL, 169–174.

[4] G. Erkan and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. of Artificial Intelligence Research* 22 (2004), 457–479.

[5] A. Farzindar and G. Lapalme. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. *Text Summarization Branches Out, Conference held in conjunction with ACL 2004* (2004), 27–38.

[6] K. Ganesan, C. Zhai, and J. Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. *Proc. COLING 2010*, 340–348.

[7] Y. Gong and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. *Proc. SIGIR 2001*, 19–25.

[8] C. Grover, B. Hachey, I. Hughson, and C. Korycinski. 2003. Automatic summarisation of legal documents. *Proc. ICAIL 2003*, 243–251.

[9] C. Grover, B. Hachey, and C. Korycinski. 2003. Summarising legal texts: Sentential tense and argumentative roles. *Proc. of the HLT-NAACL 2003 Workshop on Text summarization*, 33–40.

[10] B. Hachey and C. Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law* 14, 4 (2006), 305–345.

[11] M.-Y. Kim, Y. Xu, and R. Goebel. 2012. Summarization of legal texts with high cohesion and automatic compression rate. *JSAI International Symposium on Artificial Intelligence*, 190–204.

[12] Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. *Proc. EMNLP 2014*, 1746–1751.

[13] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 3294–3302.

[14] C.-Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out, Conference held in conjunction with ACL 2004* (2004), 74–81.

[15] A. M. Rush, S. Chopra, and J. Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *Proc. EMNLP 2015*, 379–389.

[16] J. Savelka, V. R. Walker, M. Grabmair, and K. D. Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues* 58, 2 (2017), 21–45.

[17] F. Schilder and H. Molina-Salgado. 2006. Evaluating a summarizer for legal text with a large text collection. *3rd Midwestern Computational Linguistics Colloquium (MCLC)*.

[18] C. Stab and I. Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. *Proc. EMNLP 2014*, 46–56.

[19] M. Sundararajan, A. Taly, and Q. Yan. 2017. Axiomatic Attribution for Deep Networks. *Proc. ICML 2017*, 3319–3328.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

[21] T. Vodolazova, E. Lloret, R. Muñoz, and M. Palomar. 2013. The role of statistical and semantic features in single-document extractive summarization. *Artificial Intelligence Research* 2, 3 (2013), 35.

[22] V. R. Walker, J. H. Han, X. Ni, and K. Yoseda. 2017. Semantic types for computational legal reasoning: propositional connectives and sentence roles in the veterans' claims dataset. *Proc. ICAIL 2017*, 217–226.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. *Proc. NAACL 2016*, 1480–1489.

[24] M. Yousfi-Monod, A. Farzindar, and G. Lapalme. 2010. Supervised machine learning for summarizing legal documents. *Canadian Conference on Artificial Intelligence*, 51–62.