Single Document Summarization as Tree Induction

Citation for published version:

Digital Object Identifier (DOI):
10.18653/v1/n19-1173

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics
Single Document Summarization as Tree Induction

Yang Liu, Ivan Titov and Mirella Lapata
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
yang.liu2@ed.ac.uk, ititov@inf.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

In this paper we conceptualize single-document extractive summarization as a tree induction problem. In contrast to previous approaches (Marcu, 1999; Yoshida et al., 2014) which have relied on linguistically motivated document representations to generate summaries, our model induces a multi-root dependency tree while predicting the output summary. Each root node in the tree is a summary sentence, and the subtrees attached to it are sentences whose content relates to or explains the summary sentence. We design a new iterative refinement algorithm: it induces the trees through repeatedly refining the structures predicted by previous iterations. We demonstrate experimentally on two benchmark datasets that our summarizer performs competitively against state-of-the-art methods.

1 Introduction

Single-document summarization is the task of automatically generating a shorter version of a document while retaining its most important information. The task has received much attention in the natural language processing community due to its potential for various information access applications. Examples include tools which digest textual content (e.g., news, social media, reviews), answer questions, or provide recommendations.

Of the many summarization paradigms that have been identified over the years (see Mani 2001 and Nenkova and McKeown 2011 for comprehensive overviews), two have consistently attracted attention. In abstractive summarization, various text rewriting operations generate summaries using words or phrases that were not in the original text, while extractive approaches form summaries by copying and concatenating the most important spans (usually sentences) in a document. Recent approaches to (single-document) extractive summarization frame the task as a sequence labeling problem taking advantage of the success of neural network architectures (Bahdanau et al., 2015). The idea is to predict a label for each sentence specifying whether it should be included in the summary. Existing systems mostly rely on recurrent neural networks (Hochreiter and Schmidhuber, 1997) to model the document and obtain a vector representation for each sentence (Nallapati et al., 2017; Cheng and Lapata, 2016). Inter-sentential relations are captured in a sequential manner, without taking the structure of the document into account, although the latter has been shown to correlate with what readers perceive as important in a text (Marcu, 1999). Another problem in neural-based extractive models is the lack of interpretability. While capable of identifying summary sentences, these models are not able to rationalize their predictions (e.g., a sentence is in the summary because it describes important content upon which other related sentences elaborate).

The summarization literature offers examples of models which exploit the structure of the underlying document, inspired by existing theories of discourse such as Rhetorical Structure Theory (RST; Mann and Thompson 1988). Most approaches produce summaries based on tree-like document representations obtained by a parser trained on discourse annotated corpora (Carlson et al., 2003; Prasad et al., 2008). For instance, Marcu (1999) argues that a good summary can be generated by traversing the RST discourse tree structure top-down, following nucleus nodes (discourse units in RST are characterized regarding their text importance; nuclei denote central units, whereas satellites denote peripheral ones). Other work (Hirao et al., 2013; Yoshida et al., 2014) extends this idea by transforming RST trees into dependency trees and generating summaries by tree trimming. Gerani et al. (2014) summarize product reviews; their system aggregates RST trees rep-
1. One wily coyote traveled a bit too far from home, and its resulting adventure through Harlem had alarmed residents doing a double take and scurrying to get out of its way Wednesday morning.

2. Police say frightened New Yorkers reported the coyote sighting around 9:30 a.m., and an emergency service unit was dispatched to find the animal.

3. The little troublemaker was caught and tranquilized in Trinity Cemetery on 155th street and Broadway, and then taken to the Wildlife Conservation Society at the Bronx Zoo, authorities said.


5. She said the Department of Environmental Conservation will either send the animal to a rescue center or put it back in the wild.

6. According to Adrian Benepe, New York City Parks Commissioner, coyotes in Manhattan are rare, but not unheard of.

7. “This is actually the third coyote that has been seen in the last 10 years,” Benepe said.

8. Benepe said there is a theory the coyotes make their way to the city from suburban Westchester.

9. He said they probably walk down the Amtrak rail corridor along the Hudson River or swim down the Hudson River until they get to the city.

Figure 1: Dependency discourse tree for a document from the CNN/DailyMail dataset (Hermann et al., 2015). Blue nodes indicate the roots of the tree (i.e., summary sentences) and parent-child links indicate dependency relations.

representing individual reviews into a graph, from which an abstractive summary is generated. Despite the intuitive appeal of discourse structure for the summarization task, the reliance on a parser which is both expensive to obtain (since it must be trained on labeled data) and error prone, presents a major obstacle to its widespread use.

Recognizing the merits of structure-aware representations for various NLP tasks, recent efforts have focused on learning latent structures (e.g., parse trees) while optimizing a neural network model for a down-stream task. Various methods impose structural constraints on the basic attention mechanism (Kim et al., 2017; Liu and Lapata, 2018), formulate structure learning as a reinforcement learning problem (Yogatama et al., 2017; Williams et al., 2018), or sparsify the set of possible structures (Niculae et al., 2018). Although latent structures are mostly induced for individual sentences, Liu and Lapata (2018) induce dependency-like structures for entire documents.

Drawing inspiration from this work and existing discourse-informed summarization models (Marcu, 1999; Hirao et al., 2013), we frame extractive summarization as a tree induction problem. Our model represents documents as multi-root dependency trees where each root node is a summary sentence, and the subtrees attached to it are sentences whose content is related to and covered by the summary sentence. An example of a document and its corresponding tree is shown in Figure 1; tree nodes correspond to document sentences; blue nodes represent those which should be in the summary, dependent nodes relate to or are subsumed by the parent summary sentence.

We propose a new framework that uses structured attention (Kim et al., 2017) as both the objective and attention weights for extractive summarization. Our model is trained end-to-end, it induces document-level dependency trees while predicting the output summary, and brings more interpretability in the summarization process by helping explain how document content contributes to the model’s decisions. We design a new iterative structure refinement algorithm, which learns to induce document-level structures through repeatedly refining the trees predicted by previous iterations and allows the model to infer complex trees which go beyond simple parent-child relations (Liu and Lapata, 2018; Kim et al., 2017). The idea of structure refinement is conceptually related to recently proposed models for solving iterative inference problems (Marino et al., 2018; Putzky and Welling, 2017; Lee et al., 2018). It is also related to structured prediction energy networks (Belanger et al., 2017) which approach structured prediction as iterative minimization of an energy function. However, we are not aware of any previous work considering structure refinement for tree induction problems.

Our contributions in this work are three-fold: a novel conceptualization of extractive summarization as a tree induction problem; a model which capitalizes on the notion of structured attention to learn document representations based on iterative structure refinement; and large-scale evaluation studies (both automatic and human-based) which demonstrate that our approach performs competitively against state-of-the-art methods while being able to rationalize model predictions.

2 Model Description

Let \( d \) denote a document containing several sentences \([sent_1, sent_2, \ldots, sent_m]\), where \( sent_i \) is the \( i \)-th sentence in the document. Extractive summarization can be defined as the task of assigning a label \( y_i \in \{0, 1\} \) to each \( sent_i \), indicating whether the sentence should be included in the summary. It is assumed that summary sentences represent the most important content of the document.
2.1 Baseline Model

Most extractive models frame summarization as a classification problem. Recent approaches (Zhang et al., 2018; Dong et al., 2018; Nallapati et al., 2017; Cheng and Lapata, 2016) incorporate a neural network-based encoder to build representations for sentences and apply a binary classifier over these representations to predict whether the sentences should be included in the summary. Given predicted scores $r$ and gold labels $y$, the loss function can be defined as:

$$L = - \sum_{i=1}^{m} (y_i \ln(r_i) + (1 - y_i) \ln(1 - r_i))$$  (1)$$

The encoder in extractive summarization models is usually a recurrent neural network with Long-Short Term Memory (LSTM; Hochreiter and Schmidhuber 1997) or Gated Recurrent Units (GRU; Cho et al. 2014). In this paper, our baseline encoder builds on the Transformer architecture (Vaswani et al., 2017), a recently proposed highly efficient model which has achieved state-of-the-art performance in machine translation (Vaswani et al., 2017) and question answering (Yu et al., 2018). The Transformer aims at reducing the fundamental constraint of sequential computation which underlies most architectures based on RNNs. It eliminates recurrence in favor of applying a self-attention mechanism which directly models relationships between all words in a sentence.

More formally, given a sequence of input vectors $\{x_1, x_2, \cdots, x_n\}$, the Transformer is composed of a stack of $N$ identical layers, each of which has two sub-layers:

$$\tilde{h}^l = \text{LayerNorm}(h^{l-1} + \text{MHAtt}(h^{l-1}))$$  (2)

$$h^l = \text{LayerNorm}(\tilde{h}^l + \text{FFN}(\tilde{h}^l))$$  (3)

where $h^0 = \text{PosEmb}(x)$ and $\text{PosEmb}$ is the function of adding positional embeddings to the input; the superscript $l$ indicates layer depth; $\text{LayerNorm}$ is the layer normalization operation proposed in Ba et al. (2016); $\text{MHAtt}$ represents the multi-head attention mechanism introduced in Vaswani et al. (2017) which allows the model to jointly attend to information from different representation subspaces (at different positions); and $\text{FFN}$ is a two-layer feed-forward network with ReLU as hidden activation function.

For our extractive summarization task, the baseline system is composed of a sentence-level Transformer ($\mathcal{T}_S$) and a document-level Transformer ($\mathcal{T}_D$), which have the same structure. For each sentence $s_i = [w_{i1}, w_{i2}, \cdots, w_{in}]$ in the input document, $\mathcal{T}_S$ is applied to obtain a contextual representation for each word:

$$[u_{i1}, u_{i2}, \cdots, u_{in}] = \mathcal{T}_S([w_{i1}, w_{i2}, \cdots, w_{in}])$$  (4)

And the representation of a sentence is acquired by applying weighted-pooling:

$$a_{ij} = W_0 u_{ij}^T$$  (5)

$$s_i = \frac{1}{n} \sum_{j=1}^{n} a_{ij} u_{ij}$$  (6)

Document-level transformer $\mathcal{T}_D$ takes $s_i$ as input and yields a contextual representation for each sentence:

$$[v_1, v_2, \cdots, v_m] = \mathcal{T}_D([s_1, s_2, \cdots, s_m])$$  (7)

Following previous work (Nallapati et al., 2017), we use a sigmoid function after a linear transformation to calculate the probability $r_i$ of selecting $s_i$ as a summary sentence:

$$r_i = \text{sigmoid}(W_1 v_i^T)$$  (8)

2.2 Structured Summarization Model

In the Transformer model sketched above, intersentence relations are modeled by multi-head attention based on softmax functions, which only capture shallow structural information. Our summarizer, which we call SUMO as a shorthand for Structured Summarization Model classifies sentences as summary-worthy or not, and simultaneously induces the structure of the source document as a multi-root tree. An overview of SUMO is illustrated in Figure 2. The model has the same sentence-level encoder $\mathcal{T}_S$ as the baseline Transformer model (see the bottom box in Figure 2), but differs in two important ways: (a) it uses structured attention to model the roots (i.e., summary sentences) of the underlying tree (see the upper box in Figure 2); and (b) through iterative refinement it is able to progressively infer more complex structures from past guesses (see the second and third block in Figure 2).
Algorithm 1: Calculate Tree Marginal Probabilities based on Tree-Matrix-Theorem

Function TMT(\(\tilde{r}_i\), \(\tilde{e}_{ij}\)):

\[
A_{ij} = \begin{cases} 
0 & \text{if } i = j \\
\exp(\tilde{r}_i) & \text{otherwise}
\end{cases}
\]

\[
L_{ij} = \begin{cases} 
\sum_{t'=1}^{n} A_{t' j} & \text{if } i = j \\
-A_{ij} & \text{otherwise}
\end{cases}
\]

\[
\tilde{L}_{ij} = \begin{cases} 
L_{ij} + \exp(\tilde{r}_i) & i = j \\
L_{ij} & \text{otherwise}
\end{cases}
\]

\[
e_{ij} = \begin{cases} 
(1 - \delta_{1,j}) A_{ij}[\tilde{L}^{-1}]_{jj} & \text{if } i \neq j \\
(1 - \delta_{i,1}) A_{ij}[\tilde{L}^{-1}]_{ji} & \text{otherwise}
\end{cases}
\]

\[
\hat{r}_i = \exp(\tilde{r}_i)[\tilde{L}^{-1}]_{i1}
\]

return \(r_i, e_{ij}\)

Structured Attention  Assuming document sentences have been already encoded, SUMO first calculates the unnormalized root score \(\tilde{r}_i\) for \(sent_i\) to indicate the extent to which it might be selected as root in the document tree. It also calculates the unnormalized edge score \(\tilde{e}_{ij}\) for sentence pair \(\langle sent_i, sent_j \rangle\) indicating the extent to which \(sent_i\) might be the head of \(sent_j\) in that tree (first upper block in Figure 2). To inject structural bias, SUMO normalizes these scores as the marginal probabilities of forming edges in the document dependency tree.

We use the Tree-Matrix-Theorem (TMT; Koo et al. 2007; Tutte 1984) to calculate root marginal probability \(r_i\) and edge marginal probability \(e_{ij}\), following the procedure introduced in Liu and Lapata (2017). As illustrated in Algorithm 1, we first build the Laplacian matrix \(\tilde{L}\) based on unnormalized scores and calculate marginal probabilities by matrix inverse-based operations (\(\tilde{L}^{-1}\)). We refer the interested reader to Koo et al. (2007) and Liu and Lapata (2017) for more details. In contrast to Liu and Lapata (2017), who compute the marginal probabilities of a single-root tree, our tree has multiple roots since in our task the summary typically contains multiple sentences. Given sentence vector \(s_i\) as input, SUMO computes:

\[
\tilde{r}_i = W_r s_i 
\]

\[
\tilde{e}_{ij} = s_i W_e s_j^T 
\]

\[
r_i, e_{ij} = \text{TMT}(\tilde{r}_i, \tilde{e}_{ij})
\]

Iterative Structure Refinement  SUMO essentially reduces summarization to a rooted-tree parsing problem. However, accurately predicting a tree in one shot is problematic. Firstly, when predicting the dependency tree, the model has solely access to labels for the roots (aka summary sentences), while tree edges are latent and learned without an explicit training signal. And as previous work (Liu and Lapata, 2017) has shown, a single application of TMT leads to shallow tree structures. Secondly, the calculation of \(\tilde{r}_i\) and \(\tilde{e}_{ij}\) would be based on first-order features alone, however, higher-order information pertaining to siblings and grandchildren has proved useful in discourse parsing (Carreras, 2007).

We address these issues with an inference algorithm which iteratively infers latent trees. In contrast to multi-layer neural network architectures like the Transformer or Recursive Neural Networks (Tai et al., 2015) where word representations are updated at every layer based on the output of previous layers, we refine only the tree structure during each iteration, word representations are not passed across multiple layers. Empirically, at early iterations, the model learns shallow
Algorithm 2: Structured Summarization

Model

Input: Document d
Output: Root probabilities r^K after K iterations

1. for K ← 1 to k do
   2. Calculate sentence vectors s using sentence-level Transformer T_{Ss}
   3. v^K ← s
   4. for k ← 1 to K do
      5. Calculate final unnormalized root and edge scores:
         \[ \tilde{r}_i^K = \text{TMT}(\tilde{r}_i^{K-1}) \]
         \[ \tilde{e}_{ij}^K = v_i^{K-1} W_r v_j^{K-1T} \]
      6. Calculate root probabilities:
         \[ r_i^K = \text{TMT}(\tilde{r}_i^K) \]
      7. Update sentence representations:
         \[ v_i^K = \text{k-Hop-Propagation}(e_i^K, s, k) \]
   8. end
   9. Calculate root probabilities:
      \[ r^K = \text{TMT}(\tilde{r}_i^K) \]
   10. end

Function k-Hop-Propagation(e, s, k):

11. z^K ← s
12. for l ← 1 to k do
13. \[ p_i^l = \frac{1}{n} \sum_j e_{ij} z_j^{l-1} \]
14. \[ c_i^l = \frac{1}{n} \sum_j e_{ji} z_j^{l-1} \]
15. \[ z_i^l = \tanh( W_s s_i [p_i^{l-1}, c_i^{l-1}, z_i^{l-1}] ) \]
16. end
17. return z^K

3.1 Summarization Datasets

We evaluated SUMO on two benchmark datasets, namely the CNN/DailyMail news highlights dataset (Hermann et al., 2015) and the New York Times Annotated Corpus (NYT; Sandhaus 2008). The CNN/DailyMail dataset contains news articles and associated highlights, i.e., a few bullet points giving a brief overview of the article. We used the standard splits of Hermann et al. (2015) for training, validation, and testing (90,266/1,220/1,093 CNN documents and 196,961/12,148/10,397 DailyMail documents). We did not anonymize entities.

The NYT dataset contains 110,540 articles with abstractive summaries. Following Durrett et al. (2016), we split these into 100,834 training and 9,706 test examples, based on date of publication (test is all articles published on January 1, 2007 or later). We also followed their filtering procedure, documents with summaries that are shorter than 50 words were removed from the raw dataset. The
<table>
<thead>
<tr>
<th>Model</th>
<th>CNN  R-1</th>
<th>CNN  R-2</th>
<th>CNN  R-L</th>
<th>DM   R-1</th>
<th>DM   R-2</th>
<th>DM   R-L</th>
<th>CNN+DM R-1</th>
<th>CNN+DM R-2</th>
<th>CNN+DM R-L</th>
<th>NYT  R-1</th>
<th>NYT  R-2</th>
<th>NYT  R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAD-3</td>
<td>29.2</td>
<td>11.2</td>
<td>26.0</td>
<td>40.7</td>
<td>18.3</td>
<td>37.2</td>
<td>39.6</td>
<td>17.7</td>
<td>36.2</td>
<td>35.5</td>
<td>17.3</td>
<td>32.0</td>
</tr>
<tr>
<td>Narayan et al. (2018)</td>
<td>30.4</td>
<td>11.7</td>
<td>26.9</td>
<td>41.0</td>
<td>18.8</td>
<td>37.7</td>
<td>40.0</td>
<td>18.2</td>
<td>36.6</td>
<td>41.3</td>
<td>22.0</td>
<td>37.8</td>
</tr>
<tr>
<td>Marcu (1999)</td>
<td>25.6</td>
<td>6.10</td>
<td>19.5</td>
<td>31.9</td>
<td>12.4</td>
<td>23.5</td>
<td>26.5</td>
<td>9.80</td>
<td>20.4</td>
<td>29.6</td>
<td>11.2</td>
<td>23.0</td>
</tr>
<tr>
<td>Durrett et al. (2016)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>40.8</td>
<td>22.3</td>
<td>36.7</td>
</tr>
<tr>
<td>See et al. (2017)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>39.5</td>
<td>17.3</td>
<td>36.4</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Celikyilmaz et al. (2018)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>41.7</td>
<td>19.5</td>
<td>37.9</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Transformer (no doc-att)</td>
<td>29.2</td>
<td>11.1</td>
<td>25.6</td>
<td>40.5</td>
<td>18.1</td>
<td>36.8</td>
<td>39.7</td>
<td>17.0</td>
<td>35.9</td>
<td>41.1</td>
<td>21.5</td>
<td>37.0</td>
</tr>
<tr>
<td>Transformer (1-layer doc-att)</td>
<td>29.5</td>
<td>11.4</td>
<td>26.0</td>
<td>41.5</td>
<td>18.7</td>
<td>38.0</td>
<td>40.6</td>
<td>18.1</td>
<td>36.7</td>
<td>41.8</td>
<td>22.1</td>
<td>37.8</td>
</tr>
<tr>
<td>Transformer (3-layer doc-att)</td>
<td>29.6</td>
<td>11.8</td>
<td>26.3</td>
<td>41.7</td>
<td>18.8</td>
<td>38.0</td>
<td>40.6</td>
<td>18.1</td>
<td>36.9</td>
<td>42.0</td>
<td>22.3</td>
<td>38.2</td>
</tr>
<tr>
<td>SUMO (1-layer)</td>
<td>29.5</td>
<td>11.6</td>
<td>26.2</td>
<td>41.6</td>
<td>18.8</td>
<td>37.6</td>
<td>40.5</td>
<td>18.0</td>
<td>36.8</td>
<td>42.2</td>
<td>22.1</td>
<td>38.1</td>
</tr>
<tr>
<td>SUMO (3-layer)</td>
<td>29.7</td>
<td>12.0</td>
<td>26.5</td>
<td><strong>42.0</strong></td>
<td><strong>19.1</strong></td>
<td><strong>38.0</strong></td>
<td>41.0</td>
<td>18.4</td>
<td>37.2</td>
<td>42.3</td>
<td><strong>22.7</strong></td>
<td><strong>38.6</strong></td>
</tr>
</tbody>
</table>

Table 1: Test set results on the CNN/DailyMail and NYT datasets using ROUGE $F_1$ (R-1 and R-2 are shorthands for unigram and bigram overlap, R-L is the longest common subsequence).

filtered test set includes 3,452 test examples out of the original 9,706. Compared to CNN/DailyMail, the NYT dataset contains longer and more elaborate summary sentences.

Both datasets contain abstractive gold summaries, which are not readily suited to training extractive summarization models. A greedy algorithm similar to Nallapati et al. (2017) was used to generate an oracle summary for each document. The algorithm explores different combinations of sentences and generates an oracle consisting of multiple sentences which maximize the ROUGE score with the gold summary. We assigned label 1 to sentences selected in the oracle summary and 0 otherwise and trained SUMO on this data.

### 3.2 Implementation Details

We followed the same training procedure for SUMO and various Transformer-based baselines. The vocabulary size was set to 30K. We used 300D word embeddings which were initialized randomly from $\mathcal{N}(0, 0.01)$. The sentence-level Transformer has 6 layers and the hidden size of FFN was set to 512. The number of heads in MHAtt was set to 4. Adam was used for training ($\beta_1 = 0.9$, $\beta_2 = 0.999$). We adopted the learning rate schedule from Vaswani et al. (2017) with warming-up on the first 8,000 steps. SUMO and related Transformer models produced 3-sentence summaries for each document at test time (for both CNN/DailyMail and NYT datasets).

### 3.3 Automatic Evaluation

We evaluated summarization quality using ROUGE $F_1$ (Lin, 2004). We report unigram and bigram overlap (ROUGE-1 and ROUGE-2) as a means of assessing informativeness and the longest common subsequence (ROUGE-L) as a means of assessing fluency.

Table 1 summarizes our results. We evaluated two variants of SUMO, with one and three structured-attention layers. We compared against a baseline which simply selects the first three sentences in each document (LEAD-3) and several incarnations of the basic Transformer model introduced in Section 2.1. These include a Transformer without document-level self-attention and two variants with document-level self attention instantiated with one and three layers. Several state-of-the-art models are also included in Table 1, both extractive and abstractive.

**REFRESH** (Narayan et al., 2018) is an extractive summarization system trained by globally optimizing the ROUGE metric with reinforcement learning. The system of Marcu (1999) is another extractive summarizer based on RST parsing. It uses discourse structures and RST’s notion of nuclearity to score document sentences in terms of their importance and selects the most important ones as the summary. Our re-implementation of Marcu (1999) used the parser of Zhao and Huang (2017) to obtain RST trees. Durrett et al. (2016) develop a summarization system which integrates a compression model that enforces grammaticality and coherence. See et al. (2017) present an abstractive summarization system based on...
an encoder-decoder architecture. Celikyilmaz et al.’s (2018) system is state-of-the-art in abstractive summarization using multiple agents to represent the document as well a hierarchical attention mechanism over the agents for decoding.

As far as SUMO is concerned, we observe that it outperforms a simple Transformer model without any document attention as well as variants with document attention. SUMO with three layers of structured attention overall performs best, confirming our hypothesis that document-level structure is beneficial for summarization. The results in Table 1 also reveal that SUMO and all Transformer-based models with document attention (doc-att) outperform LEAD-3 across metrics. SUMO (3-layer) is competitive or better than state-of-the-art approaches. Examples of system output are shown in Table 4.

Finally, we should point out that SUMO is superior to Marcu (1999) even though the latter employs linguistically informed document representations.

### 3.4 Human Evaluation

In addition to automatic evaluation, we also assessed system performance by eliciting human judgments. Our first evaluation quantified the degree to which summarization models retain key information from the document following a question-answering (QA) paradigm (Clarke and Lapata, 2010; Narayan et al., 2018). We created a set of questions based on the gold summary under the assumption that it highlights the most important document content. We then examined whether participants were able to answer these questions by reading system summaries alone without access to the article. The more questions a system can answer, the better it is at summarizing the document as a whole.

We randomly selected 20 documents from the CNN/DailyMail and NYT datasets, respectively and wrote multiple question-answer pairs for each gold summary. We created 71 questions in total varying from two to six questions per gold summary. We asked participants to read the summary and answer all associated questions as best they could without access to the original document or the gold summary. Examples of questions and their answers are given in Table 4. We adopted the same scoring mechanism used in Clarke and Lapata (2010), i.e., a correct answer was marked with a score of one, partially correct answers with a score of 0.5, and zero otherwise. Answers were elicited using Amazon’s Mechanical Turk platform. Participants evaluated summaries produced by the LEAD-3 baseline, our 3-layered SUMO model and multiple state-of-the-art systems. We elicited 5 responses per summary.

Table 2 (QA column) presents the results of the QA-based evaluation. Based on the summaries generated by SUMO, participants can answer 65.3% of questions correctly on CNN/DailyMail and 57.2% on NYT. Summaries produced by LEAD-3 and comparison systems fare worse, with REFRESH (Narayan et al., 2018) coming close to SUMO on CNN/DailyMail but not on NYT. Overall, we observe there is room for improvement since no system comes close to the extractive oracle, indicating that improved sentence selection would bring further performance gains to extractive approaches. Between-systems differences are all statistically significant (using a one-way ANOVA with posthoc Tukey HSD tests; \( p < 0.01 \)) with the exception of LEAD-3 and See et al. (2017) in both CNN+DM and NYT, Narayan et al. (2018) and SUMO in both CNN+DM and NYT, and LEAD-3 and Durrett et al. (2016) in NYT.

Our second evaluation study assessed the overall quality of the summaries by asking participants to rank them taking into account the following criteria: Informativeness, Fluency, and Succinctness. The study was conducted on the Amazon Mechanical Turk platform using Best-Worst Scaling (Louviere et al., 2015), a less labor-intensive alternative to paired comparisons that has been shown to produce more reliable results than rating scales (Kiritchenko and Mohammad, 2017). Participants were presented with a document and

<table>
<thead>
<tr>
<th>Model</th>
<th>CNN+DM Rank</th>
<th>CNN+DM QA</th>
<th>NYT Rank</th>
<th>NYT QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAD</td>
<td>0.07</td>
<td>40.1</td>
<td>-0.18</td>
<td>36.3</td>
</tr>
<tr>
<td>Narayan et al. (2018)</td>
<td><strong>0.21</strong></td>
<td>62.4</td>
<td>0.12</td>
<td>46.1</td>
</tr>
<tr>
<td>Durrett et al. (2016)</td>
<td>—</td>
<td>—</td>
<td>-0.11</td>
<td>40.1</td>
</tr>
<tr>
<td>See et al. (2017)</td>
<td>-0.23</td>
<td>36.6</td>
<td>-0.44</td>
<td>35.3</td>
</tr>
<tr>
<td>Celikyilmaz et al. (2018)</td>
<td>-0.64</td>
<td>37.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SUMO (3-layer)</td>
<td>0.15</td>
<td>65.3</td>
<td><strong>0.33</strong></td>
<td><strong>57.2</strong></td>
</tr>
<tr>
<td>GOLD</td>
<td>0.11</td>
<td>—</td>
<td>-0.16</td>
<td>—</td>
</tr>
<tr>
<td>ORACLE</td>
<td>0.37</td>
<td>74.6</td>
<td>0.41</td>
<td>67.1</td>
</tr>
</tbody>
</table>

Table 2: System ranking according to human judgments on summary quality and QA-based evaluation.
summaries generated from 3 out of 7 systems and were asked to decide which summary was better and which one was worse, taking into account the criteria mentioned above. We used the same 20 documents from each dataset as in our QA evaluation and elicited 5 responses per comparison. The rating of each system was computed as the percentage of times it was chosen as best minus the times it was selected as worst. Ratings range from -1 (worst) to 1 (best). As shown in Table 2 (Rank column), participants overwhelmingly prefer the extractive oracle summaries followed by SUMO and REFRESH (Narayan et al., 2018). Abstractive systems (Celikyilmaz et al., 2018; See et al., 2017; Durrett et al., 2016) perform relatively poorly in this evaluation; we suspect that humans are less forgiving to fluency errors and slightly incoherent summaries. Interestingly, gold summaries fare worse than the oracle and extractive systems. Albeit fluent, gold summaries naturally contain less detail compared to oracle-based ones; on virtue of being abstracts, they are written in a telegraphic style, often in conversational language while participants prefer the more lucid style of the extracts. All pairwise comparisons among systems are statistically significant (using a one-way ANOVA with post-hoc Tukey HSD tests; \( p < 0.01 \)) except LEAD-3 and See et al. (2017) in both CNN+DM and NTY, Narayan et al. (2018) and SUMO in both CNN+DM and NTY, and LEAD and Durrett et al. (2016) in NYT.

### 3.5 Evaluation of the Induced Structures

To gain further insight into the structures learned by SUMO, we inspected the trees it produces. Specifically, we used the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to extract the maximum spanning tree from the attention scores. We report various statistics on the characteristics of the induced trees across datasets in Table 3. We also examine the trees learned from different SUMO variants (with different numbers of iterations) in order to establish whether the iterative process yields better structures.

Specifically, we compared the dependency trees obtained from our model to those produced by a discourse parser (Zhao and Huang, 2017) trained on a corpus which combines annotations from the RST treebank (Carlson et al., 2003) and the Penn Treebank (Marcus et al., 1993). Unlike traditional RST discourse parsers (Feng and Hirst, 2014), which first segment a document into Elementary Discourse Units (EDUs) and then build a discourse tree with the EDUs as leaves, Zhao and Huang (2017) parse a document into an RST tree along with its syntax subtrees without segmenting it into EDUs. The outputs of their parser are ideally suited for comparison with our model, since we only care about document-level structures, and ignore the sub-trees within sentence boundaries. We converted the constituency RST trees obtained from the discourse parser into dependency trees using Hirao et al.’s algorithm (2013).

As can be seen in Table 3, the dependency structures induced by SUMO are simpler compared to those obtained from the discourse parser. Our trees are generally shallower, almost half of them are projective. We also calculated the percentage of head-dependency edges that are identical between learned trees and parser generated ones. Although SUMO is not exposed to any annotated trees during training, a number of edges agree with the outputs of the discourse parser. Moreover, we observe that the iterative process involving multiple structured attention layers helps generate better discourse trees. We also compare SUMO trees against a left- and right-branching baseline, where the document is trivially parsed into a left- and right-branching tree forming a chain-like structure. As shown in Table 3, SUMO outperforms these baselines (with the exception of the one-layered model on NTY). We should also point out that the edge agreement between SUMO generated trees and left/right branching trees is low (around 30% on both datasets), indicating that the trees we learn are different from a simple chain.

2EDUs roughly correspond to clauses.
A company called CyArk specializes in digital preservation of threatened ancient and historical architecture. Founded by an Iraqi-born engineer, it plans to preserve 500 World Heritage sites within five years.

What is Road Home? [the Louisiana grant program for homeowners who lost their houses to hurricanes Katrina and Rita]

When is the applicants’ deadline for the Road Home? [July 31]

Why is the program expected to cost far more than $7.5 billion? [many more families have applied than officials anticipated]

What is the shortfall projected to be? [$2.9 billion]

In 2001, the Taliban wiped out 1700 years of history in a matter of seconds, by blowing up ancient Buddha statues in central Afghanistan with dynamite. They proceeded to do so after an attempt at bringing down the 175-foot tall sculptures with anti-aircraft artillery had failed. Sadly, the event was just the first in a series of atrocities that have robbed the world of some of its most prized cultural heritage.

The Road Home, the Louisiana grant program for homeowners who lost their houses to hurricanes Katrina and Rita, is expected to cost far more than the $7.5 billion provided by the Federal Government, in part because many more families have applied than officials anticipated.

As a result, Louisiana officials on Tuesday night set a July 31 deadline for applicants, who can receive up to $150,000 to repair or rebuild their houses.

With the cutoff date, the State hopes to be able to figure out how much more money it needs to pay for the program.

The shortfall is projected to be $2.9 billion.

In 2001, the Taliban wiped out 1700 years of history in a matter of seconds, by blowing up ancient Buddha statues in central Afghanistan with dynamite. They proceeded to do so after an attempt at bringing down the 175-foot tall sculptures with anti-aircraft artillery had failed. Sadly, the event was just the first in a series of atrocities that have robbed the world of some of its most prized cultural heritage.

But historical architecture is also under threat from calamities which might well escape our control, such as earthquakes and climate change.

The thought of losing a piece of our collective history is a bleak one.

Louisiana officials set July 31 deadline for applicants, who can receive up to $150,000 to repair or rebuild their houses.

In 2001, the Taliban wiped out 1700 years of history in a matter of seconds, by blowing up ancient Buddha statues in central Afghanistan with dynamite. They proceeded to do so after an attempt at bringing down the 175-foot tall sculptures with anti-aircraft artillery had failed. Sadly, the event was just the first in a series of atrocities that have robbed the world of some of its most prized cultural heritage.

But historical architecture is also under threat from calamities which might well escape our control, such as earthquakes and climate change.

The thought of losing a piece of our collective history is a bleak one.

The Road Home, the Louisiana grant program for homeowners who lost their houses to hurricanes Katrina and Rita, is expected to cost far more than the $7.5 billion provided by the Federal Government, in part because many more families have applied than officials anticipated.

As a result, Louisiana officials on Tuesday night set a July 31 deadline for applicants, who can receive up to $150,000 to repair or rebuild their houses.

The shortfall is projected to be $2.9 billion.

4 Conclusions

In this paper we provide a new perspective on extractive summarization, conceptualizing it as a tree induction problem. We present SUMO, a Structured Summarization Model, which induces a multi-root dependency tree of a document, where roots are summary-worthy sentences, and subtrees attached to them are sentences which elaborate or explain the summary content. SUMO generates complex trees following an iterative refinement process which builds latent structures while using information learned in previous iterations. Experiments on two datasets, show that SUMO performs competitively against state-of-the-art methods and induces meaningful tree structures.

In the future, we would like to generalize SUMO to abstractive summarization (i.e., to learn latent structure for documents and sentences) and perform experiments in a weakly-supervised setting where summaries are not available but labels can be extrapolated from the article’s title or topics.

Acknowledgments

We thank Serhii Havrylov for helpful suggestions. This research is supported by a Google PhD Fellowship to the first author. We gratefully acknowledge the support of the European Research Council (Lapata, award number 681760, “Translating Multiple Modalities into Text”; Titov award number 678254, “Broad Coverage Semantic Parsing”).
References


William Thomas Tutte. 1984. Graph theory.


