CS395T: Structured Models for NLP
Lecture 22: Summarization

Greg Durrett

Administrivia

› Proposal feedback posted
› Presentation assignments posted soon

Variational Autoencoders

\[ \mathbb{E}_{q(z|x)}[\log P(x|z, \theta)] - \text{KL}(q(z|x) \| P(z)) \]

Generative model (test):

- \( z \)
- \( x \)

Autoencoder (training):

- Input \( x \)
- “inference network” \( q(z|x) \)
- Distribution over \( z \)
- Generative model
- Maximize \( P(x|z, \theta) \)

Training VAEs

For each example \( x \):

Compute \( q \) (run forward pass to compute mu and sigma)

For some number of samples:

Sample \( z \sim q \)

Compute \( P(x|z) \) and compute loss

Backpropagate to update phi, theta

Miao et al. (2015)
Autoencoders

- During training, add Gaussian noise and force the network to predict
- Same computation graph as VAE with reparameterization, add KL term to make the objective the same
- Inference network (q) is the encoder and generator is the decoder

This Lecture

- Extractive systems for multi-document summarization
- Extractive + compressive systems for single-document summarization
- Single-document summarization with neural networks

Summarization

- What makes a good summary?

Summarization

BAGHDAD/ERBIL, Iraq (Reuters) - A strong earthquake hit large parts of northern Iraq and the capital Baghdad on Sunday, and also caused damage in villages across the border in Iran where state TV said at least six people had been killed.

There were no immediate reports of casualties in Iraq after the quake, whose epicenter was in Penjwin, in Sulaimaniyah province, which is in the semi-autonomous Kurdistan region very close to the Iranian border, according to an Iraqi meteorology official.

But eight villages were damaged in Iran and at least six people were killed and many others injured in the border town of Qasr-e Shirin in Iran, Iranian state TV said.

The US Geological Survey said the quake measured a magnitude of 7.3, while an Iraqi meteorology official put its magnitude at 6.5 according to preliminary information.

Many residents in the Iraqi capital Baghdad rushed out of houses and tall buildings in panic...
Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

The quake has been felt in several Iranian cities and eight villages have been damaged. Electricity has also been disrupted at many places, suggest few TV reports.

Summary

A massive earthquake of magnitude 7.3 struck Iraq on Sunday. The epicenter was close to the Iranian border. Eight villages were damaged and six people were killed in Iran.

What makes a good summary?

Summary

A strong earthquake of magnitude 7.3 struck Iraq and Iran on Sunday. The epicenter was close to the Iranian border. Eight villages were damaged and six people were killed in Iran.

‣ Content selection: pick the right content
  ▪ Right content was repeated within and across documents
  ▪ Domain-specific (magnitude + epicenter of earthquakes are important)

‣ Generation: write the summary

‣ Extraction: pick whole sentences from the summary

‣ Compression: compress those sentences but basically just do deletion

‣ Abstraction: rewrite + reexpress content freely

Extractive Summarization: MMR

‣ Given some articles and a length budget of $k$ words, pick some sentences of total length $\leq k$ and make a summary

‣ Pick important yet diverse content: maximum marginal relevance (MMR)

While summary is $\leq k$ words

Calculate

$$MRR \triangleq \arg \max_{D_i \in R \setminus S} \left[ \lambda \text{Sim}_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} \text{Sim}_2(D_i, D_j) \right]$$

“max over all sentences not yet in the summary”

“make this sentence similar to a query”

“make this sentence maximally different from all others added so far”

Add highest MMR sentence that doesn’t overflow length

Carbonell and Goldstein (1998)
Extractive Summarization: Centroid

- Represent the documents and each sentence as bag-of-words with TF-IDF weighting

- While summary is < k words
  - Calculate score(sentence) = cosine(sent-vec, doc-vec)
  - Discard all sentences whose similarity with some sentence already in the summary is too high
  - Add the best remaining sentence that won't overflow the summary

While summary is < k words

- Calculate score(sentence) = cosine(sent-vec, doc-vec)
- Discard all sentences whose similarity with some sentence already in the summary is too high
- Add the best remaining sentence that won’t overflow the summary

Radev et al. (2004)

Extractive Summarization: Bigram Recall

- Count number of documents each bigram occurs in to measure importance

  \[
  \begin{align*}
  \text{score(massive earthquake)} &= 3 \\
  \text{score(magnitude 7.3)} &= 2 \\
  \text{score(six killed)} &= 2 \\
  \text{score(Iraqi capital)} &= 1
  \end{align*}
  \]

- Find summary that maximizes the score of bigrams it covers

- ILP formulation: \( c_{ij} \) and \( s_{ij} \) are indicator variables indexed over concepts (bigrams) and sentences, respectively

  \[
  \begin{align*}
  \text{Maximize: } & \sum_i w_i c_i \\
  \text{Subject to: } & \sum_j s_{ij} \leq L \\
  & \sum_j s_{ij} O_{ccij} \geq c_i, \quad \forall i, j \\
  & \text{“set } c_{ij} \text{ to 1 iff some sentence that contains it is included”}
  \end{align*}
  \]

- sum of included sentences’ lengths can’t exceed L

Gillick and Favre (2009)

Evaluation: ROUGE

- Rouge-n: n-gram recall of summary w.r.t. gold standard

- Rouge-2 correlates well with human judgments for multi-document summarization tasks

<table>
<thead>
<tr>
<th>Method</th>
<th>CASE</th>
<th>STEM</th>
<th>STOP</th>
<th>CASE</th>
<th>STEM</th>
<th>STOP</th>
<th>CASE</th>
<th>STEM</th>
<th>STOP</th>
<th>CASE</th>
<th>STEM</th>
<th>STOP</th>
<th>CASE</th>
<th>STEM</th>
<th>STOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-1</td>
<td>0.71</td>
<td>0.68</td>
<td>0.43</td>
<td>0.49</td>
<td>0.49</td>
<td>0.73</td>
<td>0.44</td>
<td>0.40</td>
<td>0.80</td>
<td>0.81</td>
<td>0.50</td>
<td>0.24</td>
<td>0.23</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>R-2</td>
<td>0.82</td>
<td>0.78</td>
<td>0.58</td>
<td>0.55</td>
<td>0.45</td>
<td>0.82</td>
<td>0.57</td>
<td>0.52</td>
<td>0.77</td>
<td>0.77</td>
<td>0.64</td>
<td>0.72</td>
<td>0.72</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>R-3</td>
<td>0.58</td>
<td>0.54</td>
<td>0.42</td>
<td>0.32</td>
<td>0.33</td>
<td>0.37</td>
<td>0.36</td>
<td>0.36</td>
<td>0.45</td>
<td>0.50</td>
<td>0.50</td>
<td>0.49</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>R-4</td>
<td>0.25</td>
<td>0.21</td>
<td>0.18</td>
<td>0.18</td>
<td>0.19</td>
<td>0.20</td>
<td>0.18</td>
<td>0.18</td>
<td>0.20</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>R-5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>R-6</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>R-7</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>R-8</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>R-9</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>R-10</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>R-34</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>R-93</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>R-50</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td></td>
</tr>
</tbody>
</table>

Lin (2004)

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid</td>
<td>36.03</td>
<td>7.89</td>
<td>1.20</td>
</tr>
<tr>
<td>LexRank</td>
<td>35.49</td>
<td>7.42</td>
<td>0.81</td>
</tr>
<tr>
<td>KL-Sum</td>
<td>37.63</td>
<td>8.50</td>
<td>1.26</td>
</tr>
<tr>
<td>CLASSY04</td>
<td>37.23</td>
<td>8.89</td>
<td>1.46</td>
</tr>
<tr>
<td>ICSI</td>
<td>38.02</td>
<td>9.72</td>
<td>1.72</td>
</tr>
<tr>
<td>Submodular</td>
<td>38.62</td>
<td>9.19</td>
<td>1.34</td>
</tr>
<tr>
<td>DPP</td>
<td>39.41</td>
<td>9.57</td>
<td>1.56</td>
</tr>
<tr>
<td>RegSum</td>
<td>38.23</td>
<td>9.71</td>
<td>1.59</td>
</tr>
</tbody>
</table>

Ghalandri (2017)

Caveat: these techniques all work better for multi-document than single-document!
Multi-Document vs. Single Document

- “a massive earthquake hit Iraq” “a massive earthquake struck Iraq” — lots of redundancy to help select content in multi-document case

- When you have a lot of documents, there are more possible sentences to extract:
  
  But eight villages were damaged in Iran and at least six people were killed and many others injured in the border town of Qasr-e Shirin in Iran, Iranian state TV said.

  The quake has been felt in several Iranian cities and eight villages have been damaged.

- Multi-document summarization is easier?

Compressive Summarization

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

- Sentence extraction isn’t aggressive enough at removing irrelevant content
- Want to extract sentences and also delete content from them

Syntactic Cuts

- Use syntactic rules to make certain deletions
- Delete adjuncts

A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles)...
**Syntactic Cuts**

- Use syntactic rules to make certain deletions.
- Delete second parts of coordination structures.

At least six people were killed and many others injured.

**Compressive ILP**

- Recall the Gillick+Favre ILP:
  
  \[
  \begin{align*}
  \text{Maximize:} & \quad \sum_i w_i c_i \quad \text{s.t. } s_j O_{cij} \leq c_i, \quad \forall i, j \\
  \text{Subject to:} & \quad \sum_f l_f s_j \leq L \quad \sum_j s_j O_{cij} \geq c_i, \quad \forall i
  \end{align*}
  \]

- Now \( s_j \) variables are nodes or sets of nodes in the parse tree.
- New constraint: \( s_2 \leq s_1 \)
  
  “\( s_1 \) is a prerequisite for \( s_2 \)”

At least six people were killed and many others injured.

**Compressive Summarization**

\( x_1 \) This hasn’t been Kellogg’s year.

\( x_2 \) The oat-bran craze has cost Kellogg market share.

\( x_3 \) Its president quit suddenly.

And now Kellogg is canceling its new cereal plant, which would have cost \( \$1 \) billion.

\( x_4 \)

**Constraints**

\[
\max_x \left( w^T f(x) \right) \quad \text{s.t. summary} \left( x \right) \text{ obeys length limit} \]

\[
\text{summary}(x) \text{ is grammatical} \]

\[
\text{summary}(x) \text{ is coherent} \]

**Grammaticality constraints:** allow cuts within sentences

**Coreference constraints:** do not allow pronouns that would refer to nothing

- If we’re confident about coreference, rewrite the pronoun (it → Kellogg)
- Otherwise, force its antecedent to be included in the summary

Durrett et al. (2016)
**Features**

\[
\max_x \left( w^\top f(x) \right) \quad \text{s.t.} \quad \begin{cases} \text{summary}(x) \text{ obeys length limit} \\
\text{summary}(x) \text{ is grammatical} \\
\text{summary}(x) \text{ is coherent} \end{cases}
\]

- Now uses a feature-based model, where features identify good content

\[
f(\text{And now Kellogg is canceling its new cereal plant}) = \left\{ \begin{array}{l}
\text{Centrality:} \\
\quad \mathbb{I}(\text{NumContentWords}=4) \\
\text{Document position:} \\
\quad \mathbb{I}(\text{SentenceIndex}=4) \\
\text{Lexical features:} \\
\quad \mathbb{I}(\text{FirstWord}=\text{And})
\end{array} \right.
\]

**Learning**

\[
\max_x \left( w^\top f(x) \right) \quad \text{s.t.} \quad \begin{cases} \text{summary}(x) \text{ obeys length limit} \\
\text{summary}(x) \text{ is grammatical} \\
\text{summary}(x) \text{ is coherent} \end{cases}
\]

- Train on a large corpus of New York Times documents with summaries (100,000 documents)
- Structured SVM with ROUGE as loss function
- Augment the ILP to keep track of which bigrams are included or not, use these for loss-augmented decode

Berg-Kirkpatrick et al. (2011), Durrett et al. (2016)

**Results: New York Times Corpus**

<table>
<thead>
<tr>
<th>Linguistic Quality (Human study on Mechanical Turk)</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extract clauses</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extract sentences</td>
<td>30</td>
<td>35</td>
<td>45</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First n words</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full system</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Yoshida et al. (2014)

**Neural Summarization**
Extractive paradigm isn’t all that flexible, even with compression
Training is hard! ILPs are hard! Maybe just use seq2seq?
Train to produce summary based on document

Task: generate headline from first sentence of article (can get lots of data!)

The oat bran craze has… Chopra et al. (2016)

What’s wrong with this summary?

What’s wrong with this summary?

What’s wrong with this summary?

What happens if we try this on a longer article?

Things might still go wrong, no way of preventing this…

Solutions: copy mechanism, coverage, just like in MT…

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize Nigeria’s economy. UNK says his administration is confident it will be able to thwart criminals and other Nigerians. He says the country has long endured Nigeria’s economy.

Pointer-Gen: Muhammadu Buhari says he plans to aggressively fight corruption in the northeast part of Nigeria. He says he’ll “rapidly give attention” to curbing violence in the northeast part of Nigeria. He says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: Muhammadu Buhari says he plans to aggressively fight corruption that has long plagued Nigeria. He says his administration is confident it will be able to thwart criminals. The win comes after a long history of military rule, coups and botched attempts at democracy in Africa’s most populous nation.
Neural Extractive Systems

- How abstract is this, anyway?

Challenges of Summarization

- True abstraction?
  - Not really necessary for articles
  - Generating from structured information can usually be done with templates...

Neural Systems: Results

- Copy mechanism and coverage help substantially
- Abstractive systems don’t beat a “lead” baseline on ROUGE (less n-gram overlap)

<table>
<thead>
<tr>
<th></th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>abstractive model</td>
<td>35.46</td>
</tr>
<tr>
<td>seq-to-seq + attn</td>
<td>30.49</td>
</tr>
<tr>
<td>baseline (150k vocab)</td>
<td>31.33</td>
</tr>
<tr>
<td>seq-to-seq + attn</td>
<td>36.44</td>
</tr>
<tr>
<td>baseline (50k vocab)</td>
<td></td>
</tr>
<tr>
<td>pointer-generator</td>
<td>39.53</td>
</tr>
<tr>
<td>pointer-generator + coverage</td>
<td></td>
</tr>
<tr>
<td>lead-3 baseline</td>
<td>40.34</td>
</tr>
<tr>
<td>(ours)</td>
<td></td>
</tr>
<tr>
<td>lead-3 baseline</td>
<td>39.2</td>
</tr>
<tr>
<td>(Nallapati et al., 2017)*</td>
<td></td>
</tr>
<tr>
<td>extractive model</td>
<td>39.6</td>
</tr>
<tr>
<td>(Nallapati et al., 2017)*</td>
<td></td>
</tr>
</tbody>
</table>
Challenges of Summarization

Sentence A: I saw Joe's dog, which was running in the garden.
Sentence B: The dog was chasing a cat.

Takeaways

- Extractive systems built on heuristics / ILPs work pretty well
- Compression can make things better, especially in the single-document setting
- Neural systems (like MT models) can do abstractive summarization