Learning Single-Document Summarization with Compression and Anaphoricity Constraints
Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein
{gdurrett,tberg,klein}@cs.berkeley.edu

Introduction

Why is single-document summarization harder than multi-document summarization? First, we don’t have redundancy of information as a cue. Second, we don’t have as many snippets to choose from when generating our summary. We compensate for these difficulties by learning in an end-to-end way and by giving our model flexibility in generation. Our model is trained on a large corpus (100,000 document-summary pairs) and has a rich feature set so it can learn to select important content. Sentence compression arising from syntactic and discourse parsing lets us pack more content in summaries, and constraints on anaphora ensure that critical segments of text are understandable even if they contain pronouns.

Summarization Process

Find the most relevant document content…
(extract and compress sentences)

...such that it can be rendered grammatically...
(each sentence has many possible compressions)

...with understandable pronouns.
(pronouns might need to be rendered as full mentions to be interpretable)

As the Jewel of the Seas, a luxury cruise ship owned by Royal Caribbean, steamed down the coast toward Bermuda in late October, there were plenty of distractions, including two Olympic-size pools, Latin dance lessons and Boozer Bingo.

[...]

Mr. Jaworski was one of about 150 book lovers aboard the ship for a five-day literary-themed cruise out of Boston. Known as Book It to Bermuda, it is part of a growing genre of cruises called Ship Lit. Often sponsored by publishers, the cruises feature popular authors who give readings and seminars to boatfuls of book lovers.

[...]

Model

A feature-based linear scoring function evaluates textual units (spans of sentences that serve as minimal units of extraction).

\[ \text{score}(u_i) = w^T f(u_i) \]

Sentence compression arises from syntactic and RST parses. Nested deletions yield a tree-structured set of constraints, which resemble dependencies.

For each pronoun in the document, we must either include its antecedent in the summary or rewrite the pronoun unambiguously.

For each pronoun in the document, we must either include its antecedent in the summary or rewrite the pronoun unambiguously.

\[ u_1 \rightarrow \text{coreferent} \]

Finding maximal set of words at most k words

\[ \max \sum x_i w^T f(u_i) \text{ s.t. } \sum x_i \text{len}(u_i) \leq k \]

\[ x_i = 1 \text{ indicates that } u_i \text{ is in the summary} \]

which lead to a grammatical summary

\[ s.t. \ x_i \leq x_j \text{ if } u_i \text{ depends on } u_j \]

\[ (x_i \text{ can only be included if } x_j \text{ is}) \]

Results

We evaluate on held-out New York Times data. ROUGE-2 measures content relevance; Mechanical Turk lets us evaluate clarity/grammaticality and pronoun coherence.

[Graph: Comparison of ROUGE-2 scores for different systems]

- First sentences
- Tree Knapsack (Yoshida+ 2014)
- Learned sent extractive
- Learned compressive
- Full learned system

This work

Learning and Inference

We train with structured SVM on 100,000 New York Times documents paired with abstractive summaries. We use the oracle ROUGE-1 summary within model capacity as a surrogate gold.

Subgradient for one example:

\[ \nabla \text{objective} = f(\text{oracle ROUGE sum}) - f(\text{loss-augmented sum}) \]

computed with modified ILP objective:

\[ \max \alpha \left( \sum x_i w^T f(u_i) \right) + \beta (\# \text{ unigram types recalled}) \]

(set \( \alpha = 0, \beta = 1 \) for oracle, \( \alpha = 1, \beta = -1 \) for loss-aug)

\[ \text{requires new unigram type variables and constraints} \]

Anaphoricity constraints improve pronoun handling

Clarity/grammaticality (0-10)

# unclear pronouns per summary (lower is better)