Surfer100: Generating Surveys From Web Resources, Wikipedia-style

Irene Li, Alexander Fabbri, Rina Kawamura, Yixin Liu, Xiangru Tang, Jaesung Tae, Chang Shen, Sally Ma, Tomoe Mizutani, Dragomir Radev

Introduction

Survey Generation
- A simple paragraph for introduction ❌
- A Wikipedia-style survey includes multiple sections. ✅

Problem Definition:
- Abstractive summarization from a list of related input documents;
- Generate short summary for EACH individual section.

Related work for Wikipedia page generation:
- Generating the initial leading paragraph of a Wikipedia page (Liu et al., 2018; Liu and Lapata, 2019; Perez-Beltrachini et al., 2019).

Challenges:
- No existing data: **surfer100 (100 manually written SURveys From wEb Resources on scientific topics)** for testing purposes.
- Selecting and cleaning web page: heuristics with manual checking.
- Long input sequence: two-stage method.

Contributions
- A two-stage method for generating Wikipedia-like surveys for scientific topics;
- Surfer100 dataset for survey generation using web resources.

Surfer100 Corpus

Manually selected 100 scientific topics, mainly NLP topics. For each topic:
- Web query, select top relevant html pages (input)
- Manually write summaries for each section.
- Each section: 50-150 words
- 8 annotators, each survey requires 45-60 minutes.

Download via https://github.com/Yale-LILY/Surfer100

maximum marginal relevance
perceptron
sentiment analysis
language modeling
autoencoders

gaussian mixture model
ensemble learning
lstm
gradient boosting
meta learning

Fig 4. Sample Scientific Concepts in Surfer100.

Generating Wikipedia-style surveys from web resources:
-Introduction
-History
-Main Idea
-Variations
-Applications

Fig 1. Comprehensive survey sections.

URLs
https://machinelearningmastery.com/introduction-neural-machine-translation/
https://en.wikipedia.org/wiki/Machine_translation
https://www.gala-global.org/what-machine-translation
https://www.systransoft.com/systran/translation-technology/what-is-machine-translation/
https://emerj.com/ai-sector-overviews/machine-translation-14-current-applications-and-services/

Fig 2. Sample URLs in Surfer100.

<INTRODUCTION>
Machine translation, sometimes referred to by the abbreviation MT (not to be confused with computer-aided translation, machine-aided human translation (MAHT) or interactive translation), is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. MT tools are often used to translate vast amounts of information involving millions of words that could not possibly be translated the traditional way. The quality of MT output can vary considerably; MT systems require "training" in the desired domain and language pair to increase quality.

<HISTORY>
The idea of using digital computers for translation of natural languages was proposed as early as 1946 by A. D. Booth and possibly others. Warren Weaver wrote an important memorandum "Translation" in 1949 . The Georgetown experiment was by no means the first such application, and a demonstration was made in 1954 on the APEXC machine at Birkbeck College (University of London) of a rudimentary translation of English into French. Classically, rule-based systems were used for this task, which were replaced in the 1990s with statistical methods. Since the early 2010s, a new artificial intelligence technology, deep neural networks (aka deep learning), has allowed the technology of speech recognition to reach a quality level that allowed the Microsoft Translator team to combine speech recognition with its core text translation technology to launch a new speech translation technology.
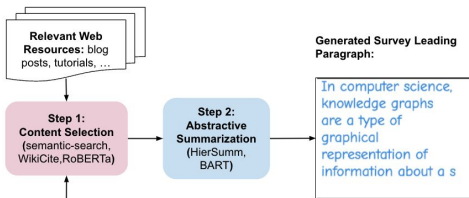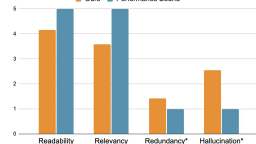
Fig 3. Sample Survey in Surfer100.

Relevant Web Resources: blog posts, tutorials, …

Generated Survey Leading Paragraph:
In computer science, knowledge graphs are a type of graphical representation of information about a s

Query Topic
"knowledge graphs"

Step 1: Content Selection (semantic-search, WikiCite,RoBERTa)

Step 2: Abstractive Summarization (HierSumm, BART)

Figure 5. Our two-stage method.

Tab 1. Human Evaluation Results.

Human Evaluation:
Randomly select 20 concepts and ask two human judges to give scores (range 1-5) on the following four perspectives: readability, relevancy, redundancy and hallucination.

Introduction
Text summarization is an interesting machine learning field that is increasingly gaining traction. As research in this area continues , we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. In this article, we look at how machine learning can be used to help shorten text.

History
Summarization has been and continues to be a hot research topic in the data science arena. While text summarization algorithms have existed for a while , major advances in natural language processing and deep learning have been made in recent years. Google has reportedly worked on projects that attempt to understand novels. Summarization can help consumers quickly understand what a book is about.

Tab 2 Sample model generated survey on the topic "text summarization".

Methodology

Step 1: Content selection
- Not every single sentence is considered to be relevant.
- Long input issue: rank all sentences with Semantic Search, WikiCite and RoBERTa-Rank.

Step 2: Abstractive Summarization
- Pre-trained models for generating abstractive summarization for each section: Hiersumm and BART.

Case Study

We show the generated summary of the topic *text summarization* in Tab 2. We could see there are descriptions about this topic: "*Text summarization is an interesting machine learning field* ", "*Automatic summarization aims to ... *". We find certain stylistic features present in the surveys do not match Wikipedia pages. For example, some content is stated in the first person: "*In this article, we...*". This is an artifact of the generation model and the content extracted and can likely be remedied by fine-tuning BART in a different setting. This aligns with Tab 1, which shows that there are hallucinations in the generated summary.

Conclusion

In this paper we show improvements in individual components of Wikipedia summarization through an application of recently-introduced embedding and summarization techniques, but largely focus on the failures of these methods when extended in a real-world scenario of full-page Wikipedia-styled summarization.

Acknowledgement