Surfer100: Generating Surveys From Web Resources on Wikipedia-style

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

Yale Department of Computer Science, USA

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

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 a
bbreviation MT (not to be confused with computer-ai
ded translation, machine-aided human translation (M
AHT) or interactive translation), is a sub-field of
 computational linguistics that investigates the us
e of software to translate text or speech from one
language to another. MT tools are often used to tra
nslate vast amounts of information involving millio
ns of words that could not possibly be translated t
he traditional way. The quality of MT output can va
ry considerably; MT systems require “training” in t
he desired domain and language pair to increase qua
lity.

<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 fir
st such application, and a demonstration was made i
n 1954 on the APEXC machine at Birkbeck College (Un
iversity of London) of a rudimentary translation of
 English into French. Classically, rule-based syste
ms were used for this task, which were replaced in
the 1990s with statistical methods. Since the early
 2010s, a new artificial intelligence technology, d
eep neural networks (aka deep learning), has allowe
d the technology of speech recognition to reach a q
uality level that allowed the Microsoft Translator
team to combine speech recognition with its core te
xt translation technology to launch a new speech tr
anslation technology.
```

Fig 3. Sample Survey in Surfer100.

- **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.

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.

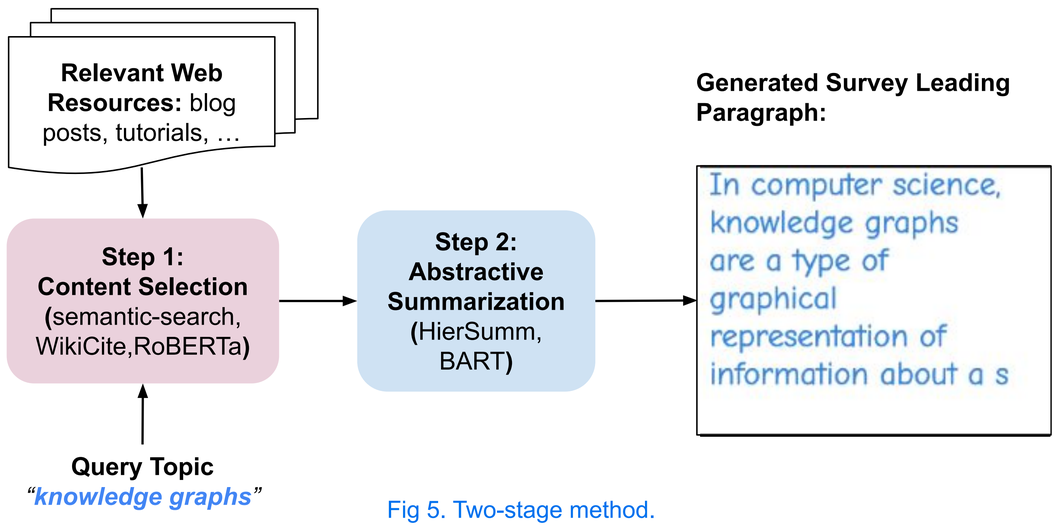
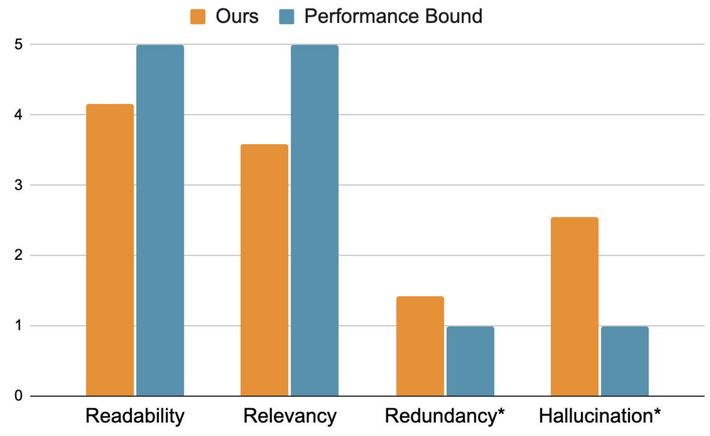


Fig 5. Two-stage method.

Evaluation



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.*

Case Study

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”.

- **Hallucination:** “in this article, …”
- **Wikipedia-style sentences:** introductory paragraph.
- Good readability and relevancy in general, but there are some hallucinations.
- The survey quality depends on the input web resources.

Conclusion

- A two-stage method for generating Wikipedia-like surveys for scientific topics;
- Surfer100 dataset for survey generation using web resources;
- Future work will be on collecting a larger corpus for survey generation, as well as a more efficient annotation pipeline;
- Advanced models for improving summarization quality.

IRENELI
ireneli.eu