BIG DATA TEXT SUMMARIZATION:

Using Deep Learning to Summarize Theses and Dissertations

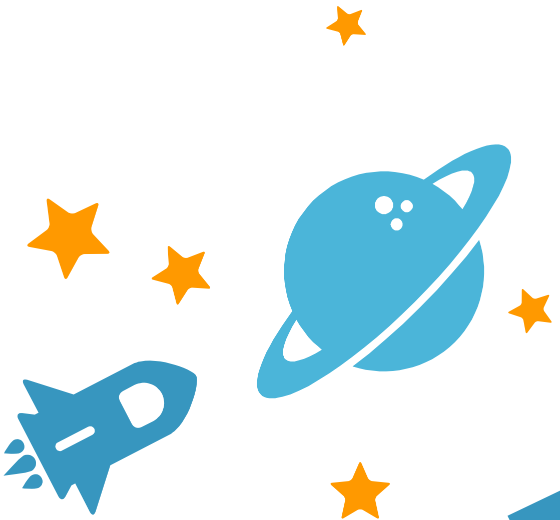
TEAM 16: Naman Ahuja, Ritesh Bansal, Bill Ingram, Palakh Jude, Sampanna Kahu, Xinyue Wang

CS4984/5984: Big Data Text Summarization, Fall 2018
Instructor: Dr. Edward A. Fox
Department of Computer Science, Virginia Tech
Blacksburg, VA 24061
December 14, 2018

OUTLINE

1. INTRODUCTION

RESEARCH QUESTIONS



This project addresses problems related to summarizing long documents by answering the following research questions.

RESEARCH QUESTIONS

RQ1

Can we identify and extract individual chapters from an ETD document?

RQ2

Can we automatically construct summaries of chapters from an ETD through existing deep learning models with non-local training data?

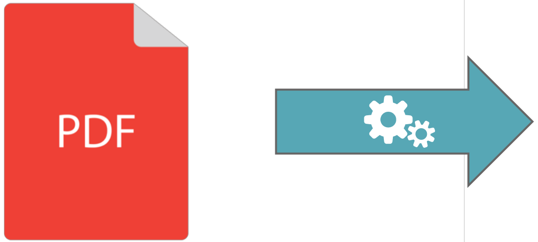
RQ3

Can we improve the quality of automatically constructed summaries for the same model through partial-local training data?

RQ4

Can we improve the quality of automatically constructed summaries for the same model through combining partial-local and non-local training data?

2. EXTRACTING ETD CHAPTERS




```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <!-- ... -->
    </teiHeader>
    <text>
        <front>
            <!-- front matter of copy text, if any, goes here -->
        </front>
        <body>
            <!-- body of copy text goes here -->
        </body>
        <back>
            <!-- back matter of copy text, if any, goes here -->
        </back>
    </text>
</TEI>
```

GROBID

GeneRation Of BIbliographic Data

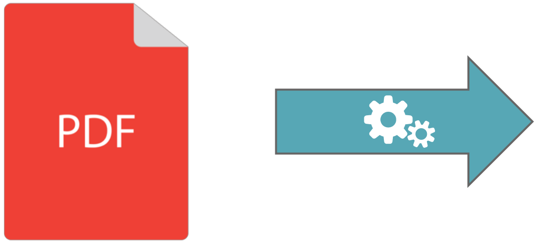
Grobid is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI-encoded documents.

CHAPTER TEXT IS EASILY EXTRACTED FROM TEI XML

```xml
<body>
    <div type="part" n="1">
        <div type="chapter" n="1">
            <head><!-- heading of part 1, chapter 1 --></head>
            <!-- text of part 1, chapter 1 -->
        </div>
        <div type="chapter" n="2">
            <!-- text of part 1, chapter 2 -->
        </div>
    </div>
    <div type="part" n="2">
        <div n="1" type="chapter">
            <!-- text of part 2, chapter 1 -->
        </div>
        <div n="2" type="chapter">
            <!-- text of part 2, chapter 2 -->
        </div>
    </div>
</body>
```

E.g., XPath:

```
//body/div[@type="chapter"]
```




```json
{
  "name": "name of the ETD goes here",
  "metadata": {
    "title": "title of the ETD goes here",
    "authors": [
      "list of authors of the ETD go here"
    ],
    "sections": [
      {
        "heading": "heading of the chapter goes here",
        "text": "text body of the chapter goes here"
      }
    ],
    "references": [
      "parsed references go here"
    ]
  }
}
```

ScienceParse

Science Parse parses scientific papers (in PDF form) and returns them into a structured JSON form.

RQ1: Can we identify and extract individual chapters from an ETD document?

- We experimented with two applications
  - Grobid
  - Science-parse

Both performed well at generating structured data from unstructured sources. Grobid was better at identifying chapters in ETDs.

3. TRAINING DATA

TRAINING DATA

CNN/DailyMail

News articles of CNN and Daily Mail and corresponding summaries.

Size: 300k records

Wikipedia All (WikiAll)

ETD related Wikipedia articles with separated introduction section and article body.

Size: 270,241 records

Wikipedia Thesis (WikiTh)

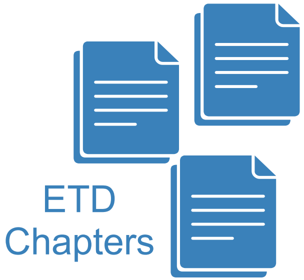
Sub-collection of Wikipedia, made up of only thesis-related Wikipedia articles.

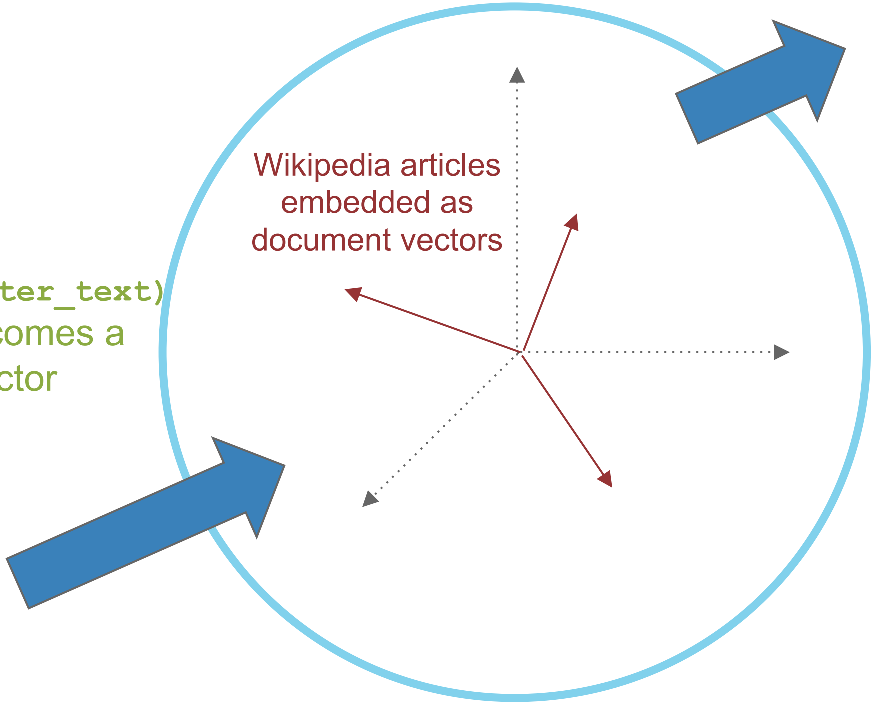
Size: 66,300 records

GET ETD RELATED WIKIPEDIA THROUGH DOC2VEC

`infer_vector(chapter_text)`

Each chapter becomes a document vector





Model trained on Wikipedia

`most_similar(chapter_vector)`

Article 1 0.876543
Article 2 0.654321
Article 3 0.987654
…

Returns vectors with shortest cosine distance

4. DEEP LEARNING MODELS

STATE-OF-THE-ART DEEP NEURAL NETWORK MODELS

- Pointer-Generator
- Sequence-to-Sequence
- Fast Abstractive Summarization-RL

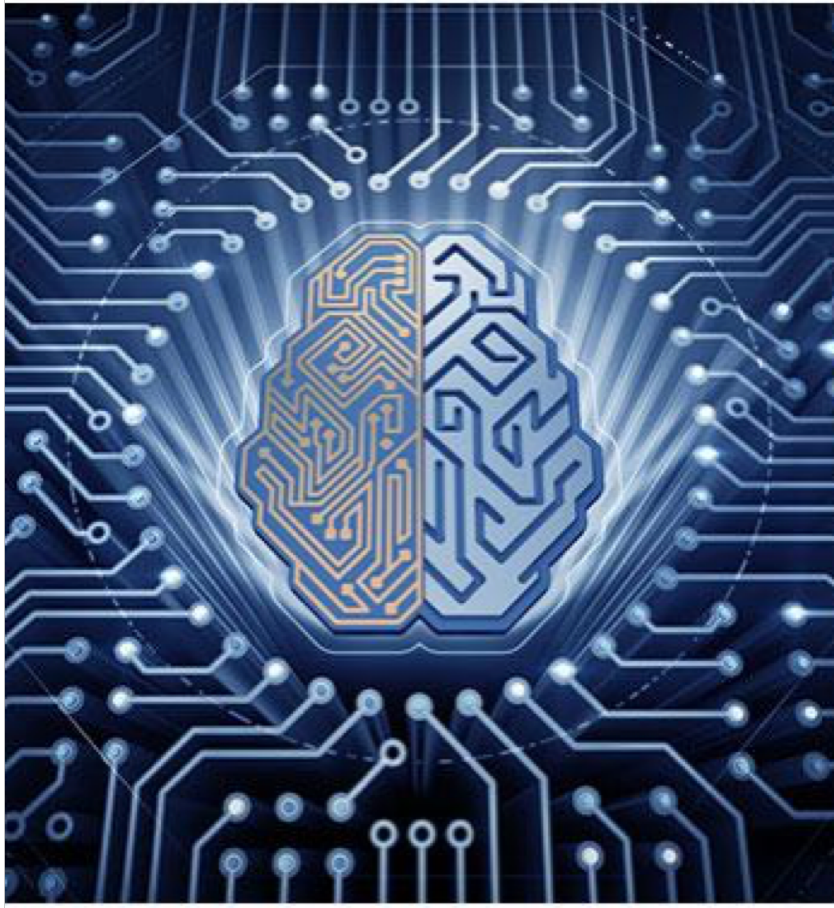


Image credit: https://www.arl.army.mil/www/default.cfm?article=3179

Training Design

We apply following training resources for each model:

- CNN DailyMail
- All ETD related Wikipedia Pages
- CNN DailyMail + Thesis related Wikipedia Pages
- CNN DailyMail + All ETD related Wikipedia Pages

Sequence-to-sequence

Hyper-parameter values used for training on CNN dataset

Max tokens of generated summary	Max source text tokens	Batch size	Vocabulary size	Number of iterations
100	400	64	40000	22K

Hyper-parameter values used for training on Wikipedia dataset

Max tokens of generated summary	Max source text tokens	Batch size	Vocabulary size	Number of iterations
500	500	64	40000	5134

Pointer-Generator

Hyper-parameter values used for training

Experiments on Hyper-parameter tuning	Max tokens of generated summary	Max source text tokens	Coverage	Batch size	Vocabulary size	Number of iterations
E1 (Finished)	100	400	NO	2	50000	50K
E2 (Finished)	500	1000	YES	2	20000	50K
E3 (In training)	500	1000	YES	4	100000	50K

Fast Abstractive Summarization - RL

Configuration used for training the abstractor and extractor models

Training Data	Vocabulary Size	Word Embedding Dimensions	Batch Size	Gradient Clipping
CNN	30000	128	32	2.0
CNN + WikiThesis	50000	300	32	2.0

Configuration used for training the RL mode (On Training)

Reward Function	Learning Rate Decay Ratio	Discount Factor	Gradient Clipping	Batch Size
rouge-l	0.5	0.95	2.0	8

5. EVALUATION

ROUGE PERFORMANCE: Sequence-to-Sequence

	1	2	L
CNN	0.20417	0.06918	0.16152
WikiAll	0.06272	0.0115	0.05816
CNN+ WikiThesis	0.01029	0.00342	0.01029
CNN+WikiAll	0.07748	0.0181	0.06899

GS: The primary source of data for this study of mentoring and instructional coaching in conjunction with new teacher support came from interviews in 4 elementary schools with four novice teachers, two mentors, and two instructional coaches. There were 3 main findings, regarding classroom procedures to academic achievement, emotional support coupled with academic and instructional guidance, and emotional support vs. academic support. Mentorship helped novice teachers learn the practical things not taught in college, and the everyday operations of a teacher. While mentorship is focused on teacher wellbeing and support, instructional coaching aims at student support and their academic success. Novice teachers indicated feelings of overwhelming isolation. They expressed feeling emotionally supported by their mentors, which helped to fight the feeling of isolation. The subjects' reported perceptions of instructional coaching were also positive. The mentors and instructional coaches reported seeing incredible value in the service they offer to novice teachers. Perceptions were learned about each program to improve their function and importance.

CNN-DailyMail: <UNK> new teachers is critical to their overall career success and ultimately influences student achievement . <UNK> <UNK> There often leave the field of education within their first five years of teaching . <UNK> <UNK> There are a variety of ways to support new teachers which are part of the induction process . <UNK>

CNN-DailyMail+WikiAll: <UNK> is a mentor of venues , such as a <UNK> teacher of the <UNK> teacher of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the nation of the <UNK> .</s>

ROUGE PERFORMANCE: Pointer-Generator

E1	1	2	L
CNN	0.23783	0.09704	0.21277
WikiAll	0.17199	0.0562	0.15392
CNN+ WikiThesis	0.19376	0.07587	0.17449
CNN+WikiAll	0.11133	0.0336	0.10251

E2	1	2	L
CNN	0.22982	0.07827	0.2068
WikiAll	0.18828	0.06504	0.17319
CNN+ WikiThesis	0.21424	0.07804	0.18966
CNN+WikiAll	0.13725	0.04079	0.1219

Applying coverage with longer encoding and decoding sequence length is improving the performance for Wikipedia training.

GS: The primary source of data for this study of mentoring and instructional coaching in conjunction with new teacher support came from interviews in 4 elementary schools with four novice teachers, two mentors, and two instructional coaches. There were 3 main findings, regarding classroom procedures to academic achievement, emotional support coupled with academic and instructional guidance, and emotional support vs. academic support. Mentorship helped novice teachers learn the practical things not taught in college, and the everyday operations of a teacher. While mentorship is focused on teacher well being and support, instructional coaching aims at student support and their academic success. Novice teachers indicated feelings of overwhelming isolation. They expressed feeling emotionally supported by their mentors, which helped to fight the feeling of isolation. The subjects' reported perceptions of instructional coaching were also positive. The mentors and instructional coaches reported seeing incredible value in the service they offer to novice teachers. Perceptions were learned about each program to improve their function and importance.

WikiAll E1:
A Novice teacher is a teacher who is interested in the field of education in the United States and Canada .
It is a part of the broader field of instructional coaching , and is a part of the broader field of instructional coaching .

CNN-DailyMail E2:
Novice shortages are occurring in the nation , but these shortages could be offset by having novice teachers stay in the field of education within their first five years of teaching .
Teachers are a variety of ways to support new teachers which are part o
the induction process .
Two are the perceptions of mentoring and instructional coaching as new elementary teacher support in NVPS ?

CNN-DailyMail + WikiTh E2:
teachers need various supports as they begin their careers in education .
Supports teachers need to recruit and hire qualified applicants .
I shortages could be offset by having novice teachers stay in the field of education longer .
The are the perceptions of mentoring and instructional coaching as new elementary teacher support in NVPS .
The are the perceptions of novice teachers and the instructional leaders -LRB- mentors and instructional coaches -RRB- toward induction program supports and what best matches the needs of a novice teacher i
public education in NVPS .
The , studying perceptions allows for the research to find how novice teachers felt toward the support and if it was what they wanted and needed as they embarked on their career .

ROUGE PERFORMANCE:
Fast Abstractive Summarization-RL

	1	2	L
CNN	0.2334	0.0939	0.1999

GS: The primary source of data for this study of mentoring and instructional coaching in conjunction with new teacher support came from interviews in 4 elementary schools with four novice teachers, two mentors, and two instructional coaches. There were 3 main findings, regarding classroom procedures to academic achievement, emotional support coupled with academic and instructional guidance, and emotional support vs. academic support. Mentorship helped novice teachers learn the practical things not taught in college, and the everyday operations of a teacher. While mentorship is focused on teacher well being and support, instructional coaching aims at student support and their academic success. Novice teachers indicated feelings of overwhelming isolation. They expressed feeling emotionally supported by their mentors, which helped to fight the feeling of isolation. The subjects' reported perceptions of instructional coaching were also positive. The mentors and instructional coaches reported seeing incredible value in the service they offer to novice teachers. Perceptions were learned about each program to improve their function and importance.

CNN-DailyMail:
novice teachers need various supports as they begin their careers.
supports exist through a variety of venues , such as a buddy teacher , a mentor , and instructional.
teachers often leave the field of education within their first five years of teaching .
this can occur with the proper support and guidance through induction programs. induction .
supporting new teachers is critical to their overall career success .

6. CONCLUSION

ROUGE PERFORMANCE

Model	Training Data	1	2	L
Seq2Seq	CNN WikiAll CNN+WikiThesis CNN+WikiAll	0.20417 0.06272 0.01029 0.07748	0.06918 0.0115 0.00342 0.0181	0.16152 0.05816 0.01029 0.0689
Pointer-Generator	CNN E2 WikiAll E2 CNN+WikiThesis E2 CNN+WikiAll E2	0.22982 0.18828 0.21424 0.13725	0.07827 0.06504 0.07804 0.04079	0.04079 0.17319 0.18966 0.1219
Fast Abstractive Summarization-RL	CNN	0.2334	0.0939	0.1999

RQ1 Can we identify and extract individual chapters from an ETD document?

Answer: Both GROBID and ScienceParse had trouble accurately parsing ETD chapters. We feel this could be improved by training them on ETDs, but we leave that for future work.

RQ2 Can we automatically construct summaries of chapters from an ETD through existing deep learning models with non-local training data?

Answer: All our models can automatically generate summaries with limited information that is related to our target

RQ3 Can we improve the quality of automatically constructed summaries for the same model through partial-local training data?

Answer: Wikipedia collection helps the model to generate different information that are related to our target comparing to RQ2

RQ4 Can we improve the quality of automatically constructed summaries for the same model through combining partial-local and non-local training data?

Answer: The model is not able to combine the good features from models trained on individual resource; The prediction of the model on ETD chapters generally fails

ACKNOWLEDGMENTS

This project was completed over the course of CS4984/5984: Big Data Text Summarization at Virginia Tech. ETD data analyzed was provided by Virginia Tech University Libraries, and the students who have submitted theses or dissertations in connection with their Virginia Tech studies.

- The authors would like to thank Dr. Edward A. Fox, Liuqing Li, and our classmates for providing insight and expertise that greatly helped us accomplish this research project.
- We thank Dr. Xuan Zhang for sharing his research on deep neural networks for user review summarization.
- We thank Dr. Srijith Rajamohan for his lecture on using Cascades and Pytorch.
- We thank Dr. Michael Horning for his informative talk on news summarization.
- We thank Yaser Keneshloo for assistance with running RLSeq2Seq and Fast Abstractive RL.
- We thank Chreston Miller for sharing his expertise with building Pointer-Generator models with TensorFlow. Although we used a different technique for our work, his in-sight was beneficial.
- We thank Frank Wanye and his team for sharing their LaTeX template with us for our report.
- We thank the other teams working on ETDs, Team 10 and Team 17, for sharing their experience, knowledge, and insights with us. We especially thank Team 10 for sharing the burden of processing the ETDs with Grobid.
- Last but not least, we are immensely grateful to the creators of all the open source software we used to create this project. Thank you very much!

THANKS!

Any questions?

TEAM 16:

Naman Ahuja, Ritesh Bansal, Bill Ingram, Palakh Jude, Sampanna Kahu, Xinyue Wang

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

» Presentation template by [SlidesCarnival]
» Photographs by [Unsplash]