Automatic Generation of Topic Labels

Areej Alokaili*
areej.okaili@sheffield.ac.uk
University of Sheffield

Nikolaos Aletras
n.aletras@sheffield.ac.uk
University of Sheffield

Mark Stevenson
mark.stevenson@sheffield.ac.uk
University of Sheffield

SIGIR 2020

M11115020 吳宇祥 M11015084 魏向晨
M11115079 張家維 M11115099 郭建鴻
M11115073 張晉嘉 M11115103 施昱民

TABLE OF CONTENTS

01

INTRODUCTION

Topic model

- Identifying the underlying themes in document collections
- An unsupervised machine learning technique
- A quick and easy way to start analyzing data
- The topic represented by a list of terms ranked by their probability can be difficult to interpret
    - Various approaches have been developed to assign descriptive labels to topics

DIFFICULTIES

- Previous work on the automatic assignment of labels to topics has relied on a two-stage approach:
  - Candidate labels are retrieved from a large pool
  - Re-ranked based on semantic similarity to the topic terms

- These extractive approaches can only assign candidate labels from a restricted set that may not include any suitable ones

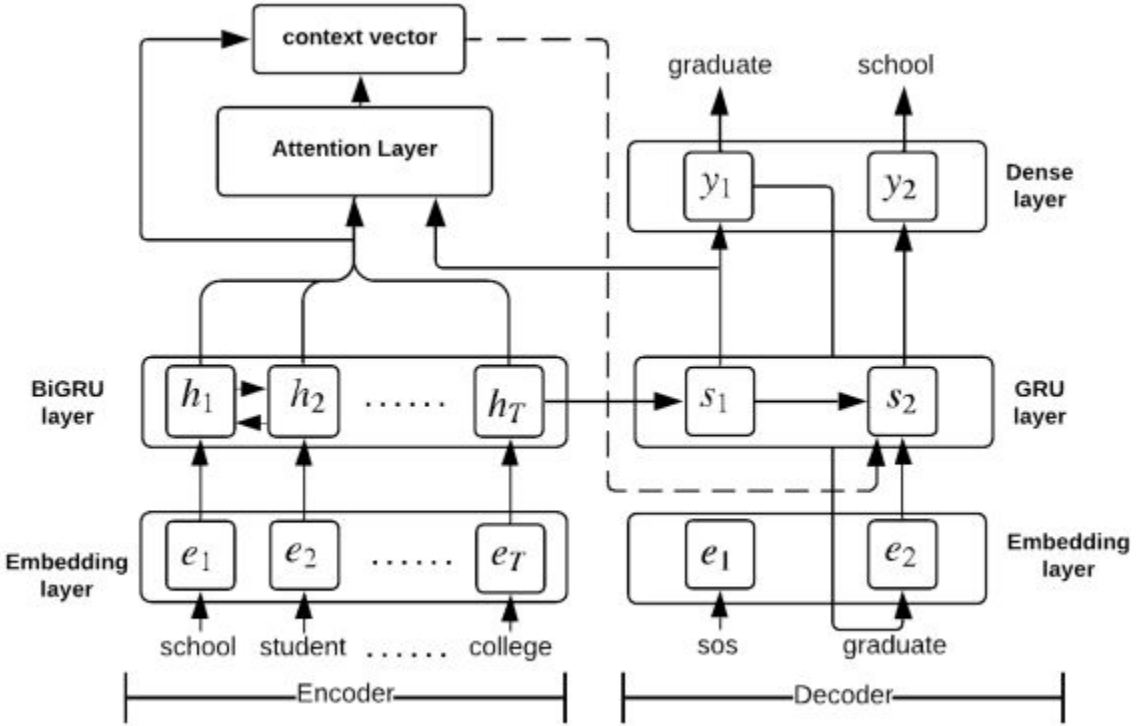
SOLUTION

- Sequence-to-sequence neural-based approach
    - The model is trained over a new large synthetic dataset created using distant supervision
    - The method is evaluated by comparing the labels which generates to ones rated by humans.

02

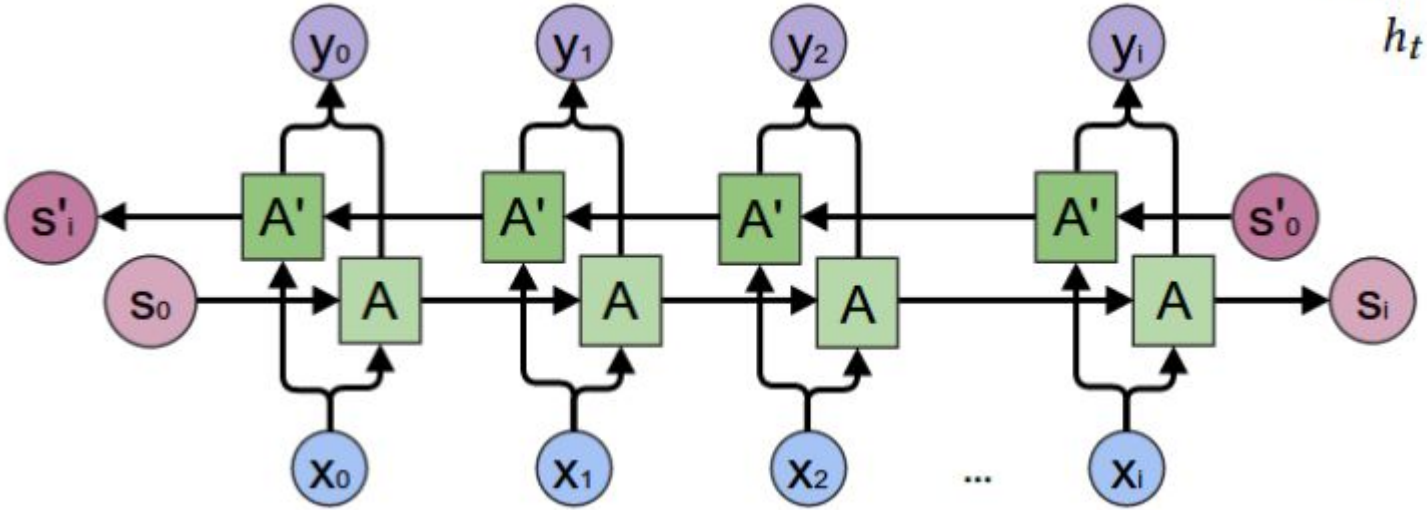
MODEL

ARCHITECTURE



BIDIRECTIONAL GRU

$$hf_t = GRU(x_t, h_{t-1})$$
$$hb_t = GRU(x_t, h_{t-1})$$
$$h_t = [hf_t; hb_t]$$



DECODER

$$c_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j$$

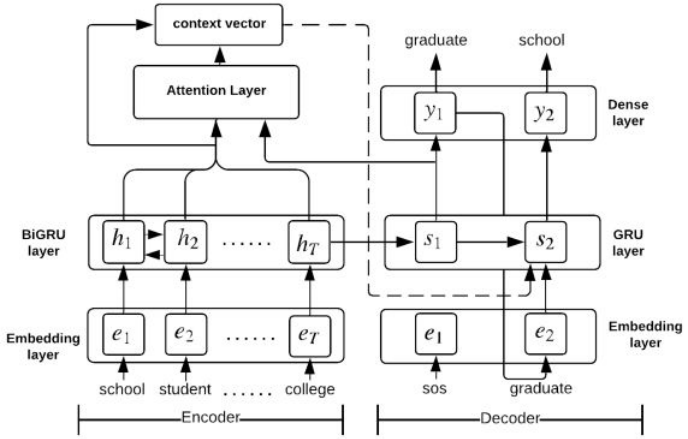
- GRU Layer :

$$s_t = GRU\left(y_{t-1}, s_{t-1}, c_t\right)$$

- Dense Layer :

$$P\left(y_t \middle| \{y_1, \cdots, y_{t-1}\} X\right) = Dense(s_t)$$

$$y_t = argmax(P\left(y_t \middle| \{y_1, \cdots, y_{t-1}\}, X\right))$$



03

Dataset

DATASET

Total 300,000 pairs of topics and labels from **ds_wiki_tfidf** & **ds_wiki_sent**

After standard preprocessing steps, which were applied to clean datasets such as removal of numbers, special characters and stop words

Total 250,506 pairs of topics and labels left

- training data:226,282(91%)
- validate data:12,424(4%)
- testing data:11,800(5%)

DATASET

Training data :

- ds_wiki_tfidf & ds_wiki_sent

Testing data :

- ds_wiki_tfidf & ds_wiki_sent
- topics_bhatia & topics_bhatia_tfidf

Automatic Labelling of Topics with Neural Embeddings

Additional testing data :

- wiki-dataset:extract from enwiki-20221120-pages-articles

zhwiki-latest-linktarget.sql.gz-rss.xml	21-Nov-2022 18:06	766
zhwiki-latest-md5sums.txt	21-Nov-2022 21:59	9278
zhwiki-latest-page.sql.gz	20-Nov-2022 16:13	243472618
zhwiki-latest-page.sql.gz-rss.xml	21-Nov-2022 18:07	748
zhwiki-latest-page_props.sql.gz	20-Nov-2022 16:14	40657538
zhwiki-latest-page_props.sql.gz-rss.xml	21-Nov-2022 18:07	766
zhwiki-latest-page_restrictions.sql.gz	20-Nov-2022 16:14	519176
zhwiki-latest-page_restrictions.sql.gz-rss.xml	21-Nov-2022 18:07	787
zhwiki-latest-pagelinks.sql.gz	20-Nov-2022 16:03	1121341240
zhwiki-latest-pagelinks.sql.gz-rss.xml	21-Nov-2022 18:06	763
zhwiki-latest-pages-articles-multistream-index...>	21-Nov-2022 02:47	35378851
zhwiki-latest-pages-articles-multistream-index...>	21-Nov-2022 21:59	835
zhwiki-latest-pages-articles-multistream-index1..>	21-Nov-2022 02:20	702484
zhwiki-latest-pages-articles-multistream-index1..>	21-Nov-2022 21:59	868
zhwiki-latest-pages-articles-multistream-index2..>	21-Nov-2022 02:21	1913845
zhwiki-latest-pages-articles-multistream-index2..>	21-Nov-2022 21:59	883
zhwiki-latest-pages-articles-multistream-index3..>	21-Nov-2022 02:23	3867086
zhwiki-latest-pages-articles-multistream-index3..>	21-Nov-2022 21:59	886
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 02:26	6279169
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 02:21	2423050
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 02:23	4631468
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 02:29	2474806
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 02:35	6449017
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index6..>	04-Nov-2022 11:20	6491030
zhwiki-latest-pages-articles-multistream-index6..>	07-Nov-2022 02:12	889
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 02:34	6598255
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream.xml.bz2	21-Nov-2022 02:47	2694577747
zhwiki-latest-pages-articles-multistream.xml.bz..>	21-Nov-2022 21:59	817
zhwiki-latest-pages-articles-multistream1.xml-p..>	21-Nov-2022 02:20	222027702
zhwiki-latest-pages-articles-multistream1.xml-p..>	21-Nov-2022 21:59	850
zhwiki-latest-pages-articles-multistream2.xml-p..>	21-Nov-2022 02:21	281203796
zhwiki-latest-pages-articles-multistream2.xml-p..>	21-Nov-2022 21:59	865
zhwiki-latest-pages-articles-multistream3.xml-p..>	21-Nov-2022 02:23	366642855
zhwiki-latest-pages-articles-multistream3.xml-p..>	21-Nov-2022 21:59	868
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 02:26	291756311
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 02:21	138309890
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 02:23	318977219
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 02:29	198039574
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream6.xml-p..>	21-Nov-2022 02:35	473913869
zhwiki-latest-pages-articles-multistream6.xml-p..>	21-Nov-2022 21:59	871

04

Experiment

EVALUATION

$$score\_topic_t = \max_{i=[1,\ldots,n]} BERTScore(l_t, gold\_l_{ti})$$

$$score\_model = \frac{1}{T}\sum_{t=1}^{T} score\_topic_t$$

BERTScore: Evaluating Text Generation with BERT (ArXiv 2019)

EXPERIMENT

				BERTScore		
				P	**R**	**F**
	Baselines		Top-2 label	0.902	0.912	0.902
			Top-3 label	0.870	0.903	0.882
Train data	ds_wiki_tfidf	**Test data**	topics_bhatia	0.922*†	0.928*†	0.922*†
			topics_bhatia_tfidf	0.926*†	0.930*†	0.925*†
	ds_wiki_sent		topics_bhatia	0.919†	0.926†	0.919†
			topics_bhatia_tfidf	**0.930***†	**0.933***†	**0.929***†

* and † denote statistically significant difference ($p < 0.001$) compared to Top-2 label and Top-3 label, respectively.

Our result	ds_wiki_tfidf	Another wiki-data	0.916	0.868	0.889

DATASET EXTRACT

https//dumps.wikimpedia.org/zhwiki/latest/

```
from gensim.corpora.wikicorpus import extract_pages,filter_wiki
import bz2file
import re
import opencc
from tqdm import tqdm
import codecs
wiki = extract_pages(bz2file.open('enwiki-20221120-pages-articles-multistream19.xml

def wiki_replace(d):
    s = d[1]
    s = re.sub(':*{\|[\s\S]*?\|}', '', s)
    s = re.sub('<gallery>[\s\S]*?</gallery>', '', s)
    s = re.sub('(.){{{[^{}\n]*?\|[^{}\n]*?)}}', '\\1[[\\2]]', s)
    s = filter_wiki(s)
    s = re.sub('\* *\n|\'{2,}', '', s)
    s = re.sub('\n+', '\n', s)
    s = re.sub('\n[:;]|\n +', '\n', s)
    s = re.sub('\n==', '\n\n==', s)
    s = u'【' + d[0] + u'】\n' + s
    print(s)
    return s.strip()

i = 0
f = codecs.open('wiki.txt', 'w', encoding='utf-8')
w = tqdm(wiki, desc=u'已獲取篇文章')
for d in w:
    if not re.findall('^[a-zA-Z]+:', d[0]) and d[0] and not re.findall(u'^#', d[1]):
        s = wiki_replace(d)
        f.write(s+'\n\n\n')
        i += 1
        if i % 100 == 0:
            w.set_description(u'已獲取%s篇文章'%i)

f.close()
```

zhwiki-latest-linktarget.sql.gz-rss.xml	21-Nov-2022 18:06	766
zhwiki-latest-md5sums.txt	21-Nov-2022 21:59	9278
zhwiki-latest-page.sql.gz	20-Nov-2022 16:13	243472618
zhwiki-latest-page.sql.gz-rss.xml	21-Nov-2022 18:07	748
zhwiki-latest-page_props.sql.gz	20-Nov-2022 16:14	40657538
zhwiki-latest-page_props.sql.gz-rss.xml	21-Nov-2022 18:07	766
zhwiki-latest-page_restrictions.sql.gz	20-Nov-2022 16:14	519176
zhwiki-latest-page_restrictions.sql.gz-rss.xml	21-Nov-2022 18:07	787
zhwiki-latest-pagelinks.sql.gz	20-Nov-2022 16:03	1121341240
zhwiki-latest-pagelinks.sql.gz-rss.xml	21-Nov-2022 18:06	763
zhwiki-latest-pages-articles-multistream-index...>	21-Nov-2022 02:47	35378851
zhwiki-latest-pages-articles-multistream-index...>	21-Nov-2022 21:59	835
zhwiki-latest-pages-articles-multistream-index1..>	21-Nov-2022 02:20	702484
zhwiki-latest-pages-articles-multistream-index1..>	21-Nov-2022 21:59	868
zhwiki-latest-pages-articles-multistream-index2..>	21-Nov-2022 02:21	1913845
zhwiki-latest-pages-articles-multistream-index2..>	21-Nov-2022 21:59	883
zhwiki-latest-pages-articles-multistream-index3..>	21-Nov-2022 02:23	3867086
zhwiki-latest-pages-articles-multistream-index3..>	21-Nov-2022 21:59	886
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 02:26	6279169
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 02:21	2423050
zhwiki-latest-pages-articles-multistream-index4..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 02:23	4631468
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 02:29	2474806
zhwiki-latest-pages-articles-multistream-index5..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 02:35	6449017
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream-index6..>	04-Nov-2022 11:20	6491030
zhwiki-latest-pages-articles-multistream-index6..>	07-Nov-2022 02:12	889
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 02:34	6598255
zhwiki-latest-pages-articles-multistream-index6..>	21-Nov-2022 21:59	889
zhwiki-latest-pages-articles-multistream.xml.bz2	21-Nov-2022 02:47	2694577747
zhwiki-latest-pages-articles-multistream.xml.bz..>	21-Nov-2022 21:59	817
zhwiki-latest-pages-articles-multistream1.xml-p..>	21-Nov-2022 02:20	222027702
zhwiki-latest-pages-articles-multistream1.xml-p..>	21-Nov-2022 21:59	850
zhwiki-latest-pages-articles-multistream2.xml-p..>	21-Nov-2022 02:21	281203796
zhwiki-latest-pages-articles-multistream2.xml-p..>	21-Nov-2022 21:59	865
zhwiki-latest-pages-articles-multistream3.xml-p..>	21-Nov-2022 02:23	366642855
zhwiki-latest-pages-articles-multistream3.xml-p..>	21-Nov-2022 21:59	868
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 02:26	291756311
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 02:21	138309890
zhwiki-latest-pages-articles-multistream4.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 02:23	318977219
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 02:29	198039574
zhwiki-latest-pages-articles-multistream5.xml-p..>	21-Nov-2022 21:59	871
zhwiki-latest-pages-articles-multistream6.xml-p..>	21-Nov-2022 02:35	473913869
zhwiki-latest-pages-articles-multistream6.xml-p..>	21-Nov-2022 21:59	871

CASE STUDY ANOTHER WIKI DATA

https://dumps.wikimedia.org/enwiki/latest/

Original data(term) : vmware server virtual oracle update virtualization application infrastructure management microsoft plesk web hosting dns windows linux accounts reseller software aps sql administrator panel cloudlinux automation swsoft platform packaging designed versions

Ground truth : cloud computing, microsoft exchange server, vmware, web application, virtualization, operating system

Predict : vmware server

Conclusion

- Present the first seq2seq model to generate textual labels for automatically generated topics.

- Presented a dataset built from Wikipedia and use BERTScore to measure the similarities between the generated labels and gold standard labels.

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

實作心得

- Present the first seq2seq model to generate textual labels for automatically generated topics.

- Presented a dataset built from Wikipedia and use BERTScore to measure the similarities between the generated labels and gold standard labels.

實作心得

- Present the first seq2seq model to generate textual labels for automatically generated topics.

- Presented a dataset built from Wikipedia and use BERTScore to measure the similarities between the generated labels and gold standard labels.