INTRO TO NLP - CS7.401.S23

Extracting Keyphrases and Relations from Scientific Publications

Team-62
2022201030-Amitabh Paliwal
2022201035-Prateek Pandey

- Extracting keyphrases and relations from scientific publications is the process of automatically identifying important words or phrases and their relationships to summarize the main topics or concepts discussed in a publication.
- This task is crucial as it helps researchers and readers to quickly understand the content and contributions of a publication, and enables more efficient information retrieval and literature search

Datasets

There are three benchmark datasets commonly used for automatic keyphrase extraction from scientific articles.

- The SemEval-2010 Task 5 dataset contains annotated keyphrases for 244 scientific articles from various domains.
- The ScienceIE 2017 dataset contains 500 journal articles evenly distributed among three domains and includes plain text, annotated standoff documents, and original full article text.
- The Inspec dataset contains 2,000 abstracts of scientific journal papers from Computer Science, each with two sets of keywords assigned: controlled keywords and uncontrolled keywords. These datasets are used to evaluate the performance of keyphrase extraction models.

Preprocessing

- We have used a preprocessed Inpec Dataset from Hugging face
- For Semval 2010/2017 we have used preprocessed dataset from IIITD’s Midas lab’s dataset

Models Used

Statistical Models

TFIDF

- To extract keyphrases, the team employed TF-IDF transformers which involved converting the input text into multiple N-grams ranging from 1 to n. we chose n=3.
- The text data was then converted into numbers using a Count Vectorizer.
- For each feature in the text, the TF-IDF frequency was calculated using the TF-IDF transformers. The top 15 features with the highest TF-IDF frequency were then extracted.
- With Accuracy < 0.1

Neural Models

Bilstm-CRF-We Tried Bilstm-CRF with multiple embeddings(with glove/without glove) and datasets like

- **SemEval-2010**
- SemEval-2017
- Inspec

Architecture

```
BiLSTMCRF(
  (sent_vocab): Vocab()
  (tag_vocab): Vocab()
  (embedding): Embedding(3937, 300)
  (dropout): Dropout(p=0.5, inplace=False)
  (encoder): LSTM(300, 300, bidirectional=True)
  (hidden2emit_score): Linear(in_features=600, out_features=5, bias=True)
)
```

Semval 2017

Without Glove

	precision	recall	f1-score	support
2	0.18	0.23	0.20	625
3	0.01	0.53	0.02	15
4	0.06	0.10	0.07	378
5	0.28	0.25	0.26	1394
6	0.94	0.85	0.89	19591
7	0.00	0.00	0.00	0
8	0.01	0.21	0.01	28
accuracy			0.78	22031
macro avg	0.21	0.31	0.21	22031
weighted avg	0.86	0.78	0.82	22031

Semval 2017

With Glove

	precision	recall	f1-score	support
2	0.00	0.33	0.01	6
3	0.00	0.00	0.00	0
4	0.09	0.19	0.12	378
5	0.01	0.23	0.02	31
6	0.23	0.23	0.23	1228
7	0.00	0.00	0.00	0
8	0.96	0.83	0.89	20388
accuracy			0.79	22031
macro avg	0.19	0.26	0.18	22031
weighted avg	0.91	0.79	0.84	22031

Semval 2010

Without Glove

	precision	recall	f1-score	support
2	0.23	0.33	0.27	506
3	0.18	0.30	0.23	610
4	0.97	0.94	0.95	19553
accuracy			0.90	20669
macro avg	0.46	0.52	0.48	20669
weighted avg	0.92	0.90	0.91	20669

Semval 2010

With Glove

	precision	recall	f1-score	support
2	0.07	0.48	0.12	104
3	0.05	0.47	0.09	105
4	1.00	0.92	0.96	20460
accuracy			0.92	20669
macro avg	0.37	0.62	0.39	20669
weighted avg	0.99	0.92	0.95	20669

Inspec

Without Glove

	precision	recall	f1-score	support
2	0.29	0.56	0.38	2532
3	0.95	0.90	0.92	60313
4	0.44	0.56	0.49	4447
accuracy			0.86	67292
macro avg	0.56	0.67	0.60	67292
weighted avg	0.89	0.86	0.88	67292

Inspec

With Glove

	precision	recall	f1-score	support
2	0.49	0.55	0.52	5042
3	0.34	0.54	0.42	3034
4	0.94	0.90	0.92	59216
accuracy			0.86	67292
macro avg	0.59	0.66	0.62	67292
weighted avg	0.88	0.86	0.87	67292

BiLstm with Self Attention(Overfits)

Model Tends to Overfit

	precision	recall	f1-score	support
0	0.87	0.95	0.91	123449
1	0.16	0.00	0.01	9633
2	0.07	0.06	0.06	8426
accuracy			0.84	141508
macro avg	0.37	0.34	0.33	141508
weighted avg	0.78	0.84	0.80	141508

```
attbilstm(
  (positionalencoding): PositionalEmbedding()
  (embedding): Embedding(13832, 256)
  (attention): MultiHeadAttention(
    (query_matrix): Linear(in_features=64, out_features=64, bias=False)
    (key_matrix): Linear(in_features=64, out_features=64, bias=False)
    (value_matrix): Linear(in_features=64, out_features=64, bias=False)
    (out): Linear(in_features=512, out_features=512, bias=True)
  )
  (encoder): LSTM(256, 256, num_layers=2, dropout=0.5, bidirectional=True)
  (fc): Linear(in_features=512, out_features=3, bias=True)
  (dropout): Dropout(p=0.5, inplace=False)
)
```

Bilstm(Seq2Seq) with attention(Overfits)

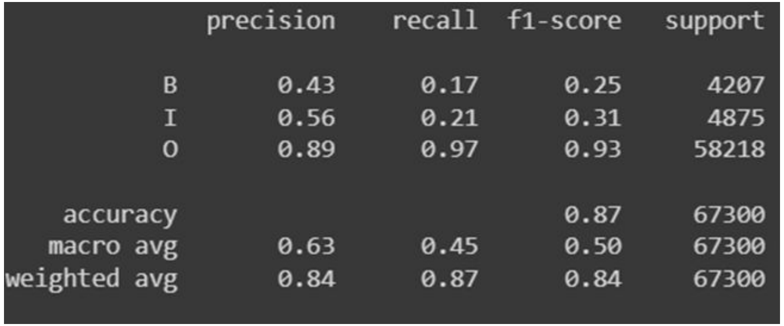
```
Encoder(
  (embedding): Embedding(13832, 128)
  (lstm): LSTM(128, 128, bidirectional=True)
  (lstm_2): LSTM(384, 128, bidirectional=True)
  (out): Linear(in_features=256, out_features=3, bias=True)
)
```

	precision	recall	f1-score	support
0	0.87	1.00	0.93	123449
1	0.00	0.00	0.00	9633
2	0.00	0.00	0.00	8426
accuracy			0.87	141508
macro avg	0.29	0.33	0.31	141508
weighted avg	0.76	0.87	0.81	141508

Pretrained Models

Keybert with keyphrase-vectorizers

KeyphraseVectorizers package together with KeyBERT. The KeyphraseVectorizers package extracts keyphrases with part-of-speech patterns from a collection of text documents and converts them into a document-keyphrase matrix.



	precision	recall	f1-score	support
B	0.43	0.17	0.25	4207
I	0.56	0.21	0.31	4875
O	0.89	0.97	0.93	58218
accuracy			0.87	67300
macro avg	0.63	0.45	0.50	67300
weighted avg	0.84	0.87	0.84	67300

Keyphrase-generation-keybart

This model uses KeyBART as its base model and fine-tunes it on the Inspec dataset

	precision	recall	f1-score	support
B	0.56	0.38	0.45	4207
I	0.63	0.41	0.50	4875
O	0.91	0.96	0.94	58218
accuracy			0.88	67300
macro avg	0.70	0.58	0.63	67300
weighted avg	0.87	0.88	0.87	67300

Keyphrase-extraction-kbir

This model uses KBIR as its base model and fine-tunes it on the Inspec dataset. KBIR or Keyphrase Boundary Infilling with Replacement is a pre-trained model which utilizes a multi-task learning setup for optimizing a combined loss of Masked Language Modeling (MLM), Keyphrase Boundary Infilling (KBI) and Keyphrase Replacement Classification (KRC)

	precision	recall	f1-score	support
B	0.63	0.65	0.64	4207
I	0.70	0.71	0.70	4875
O	0.95	0.95	0.95	58218
accuracy			0.91	67300
macro avg	0.76	0.77	0.76	67300
weighted avg	0.91	0.91	0.91	67300

Keyphrase-extraction-distilbert

This model uses distilbert as its base model and fine-tunes it on the Inspec dataset.

	precision	recall	f1-score	support
B	0.51	0.68	0.58	4207
I	0.62	0.70	0.66	4875
O	0.96	0.92	0.94	58218
accuracy			0.89	67300
macro avg	0.70	0.77	0.73	67300
weighted avg	0.90	0.89	0.90	67300

Keyphrase-generation-t5-small

This model uses T5-small model as its base model and fine-tunes it on the Inspec dataset. Keyphrase generation transformers are fine-tuned as a text-to-text generation problem where the keyphrases are generated

	precision	recall	f1-score	support
B	0.40	0.34	0.37	4207
I	0.62	0.36	0.46	4875
O	0.91	0.95	0.93	58218
accuracy			0.87	67300
macro avg	0.64	0.55	0.58	67300
weighted avg	0.85	0.87	0.86	67300

THANK YOU