Automated Essay Scoring, the Current State-Of-The-Art(SOTA) and the Future

Yongchao Wu
yongchao.wu@dsv.su.se

Stockholm University

Automated Essay Scoring

Automated Essay Scoring (AES) automatically allocates scores to essays at scale.

Manually Essay Scoring

Student essays  →  Rater → Essays Scores

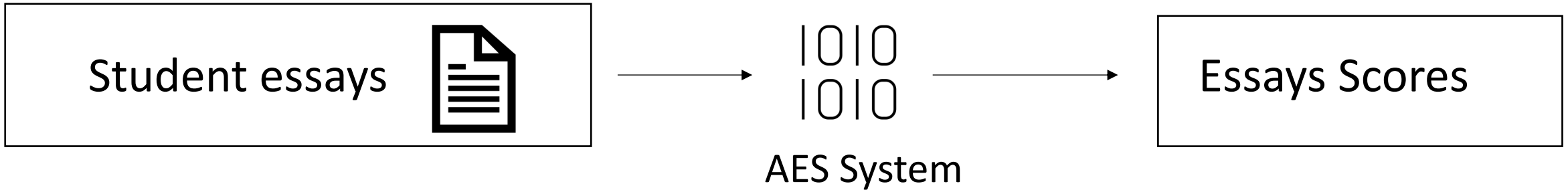
Ideally:

- Organisation
- Level of content
- Development
- Grammer & Mechanic
- Style
- Format

In fact:

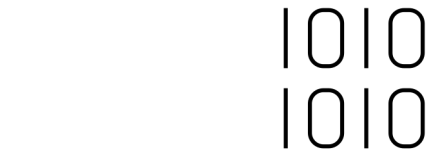
- Rater's mode
- Rater's energy
- Rater's sense of responsibility
- …
- Inter/intra rater inconsistency

Automated Essay Scoring



Ruthless
Tireless
Automatic
Consistent

Automated Essay Scoring



AES System

Essays Scores

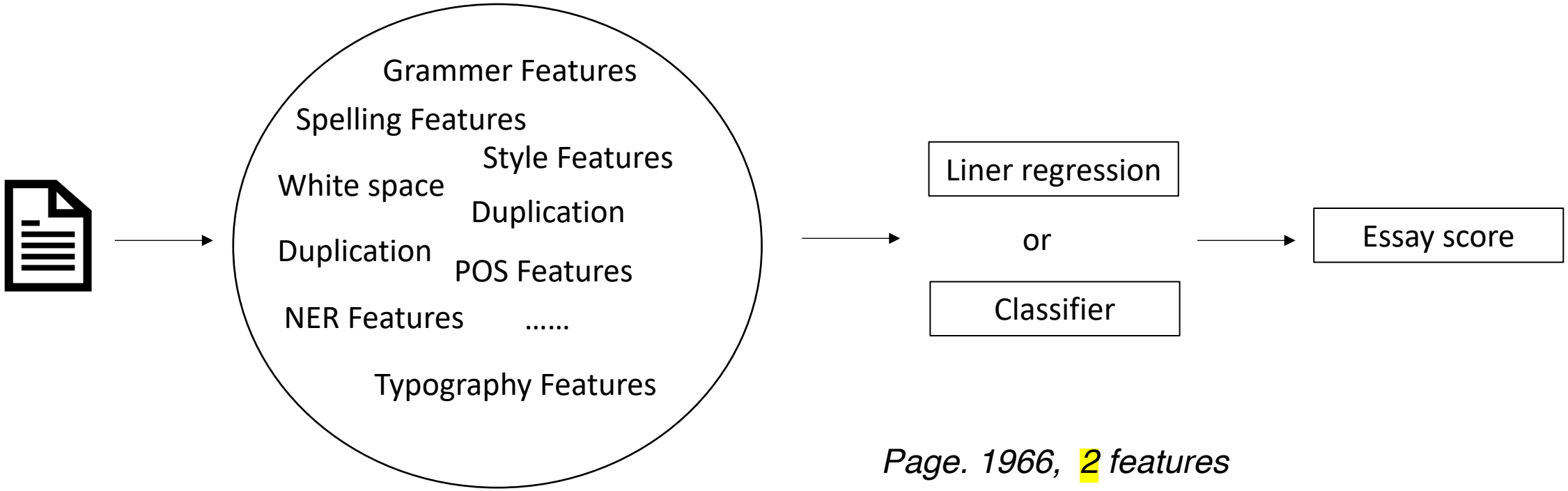
Two lines of work pursuiting SOTA

- ➢ Feature-based methods
- ➢ Neural-based methods

Translate AES into NLP tasks

- ✓ Classification Task (A, B, C, D)
- ✓ Regression Task (95, 85, 60, 30)

Feature-based AES systems



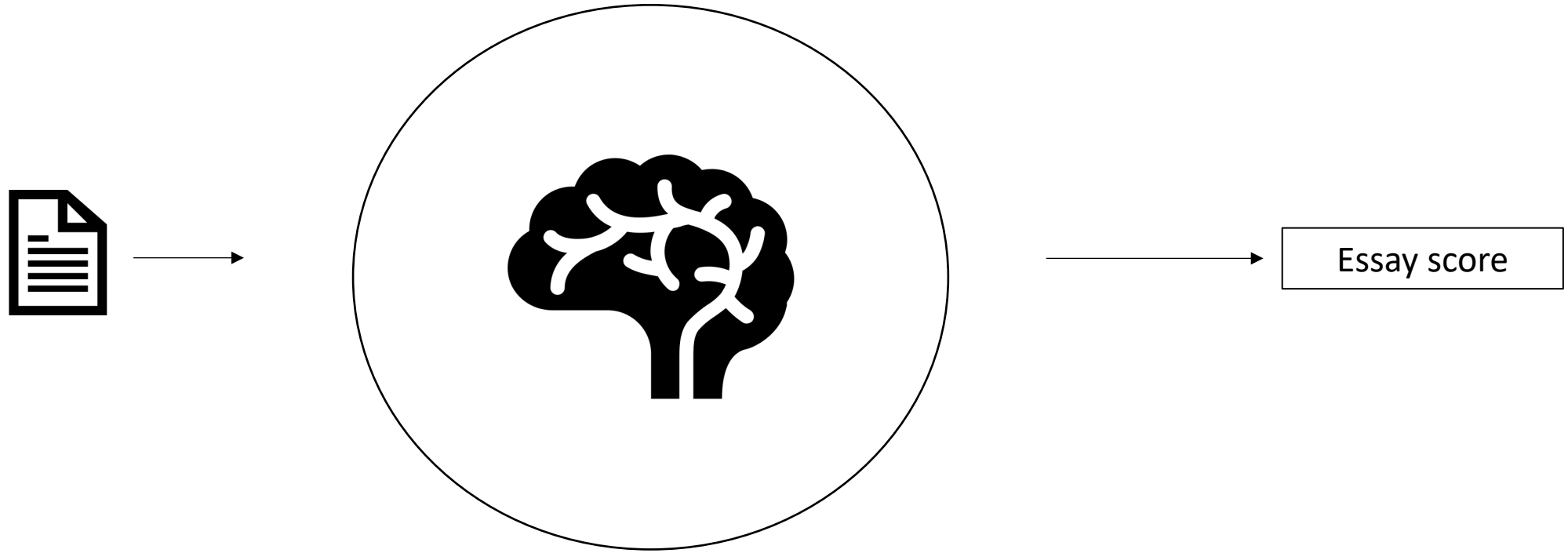
Feature Enginnering

Page. 1966, 2 features

E-rater. 2004, 10 features

Kumar et al. 2021, 1592 features

Neural-based AES systems



Nerual Architecture Engineering

Current AES Benchmark

Stockholm University

Table 1. ASAP dataset statistics.

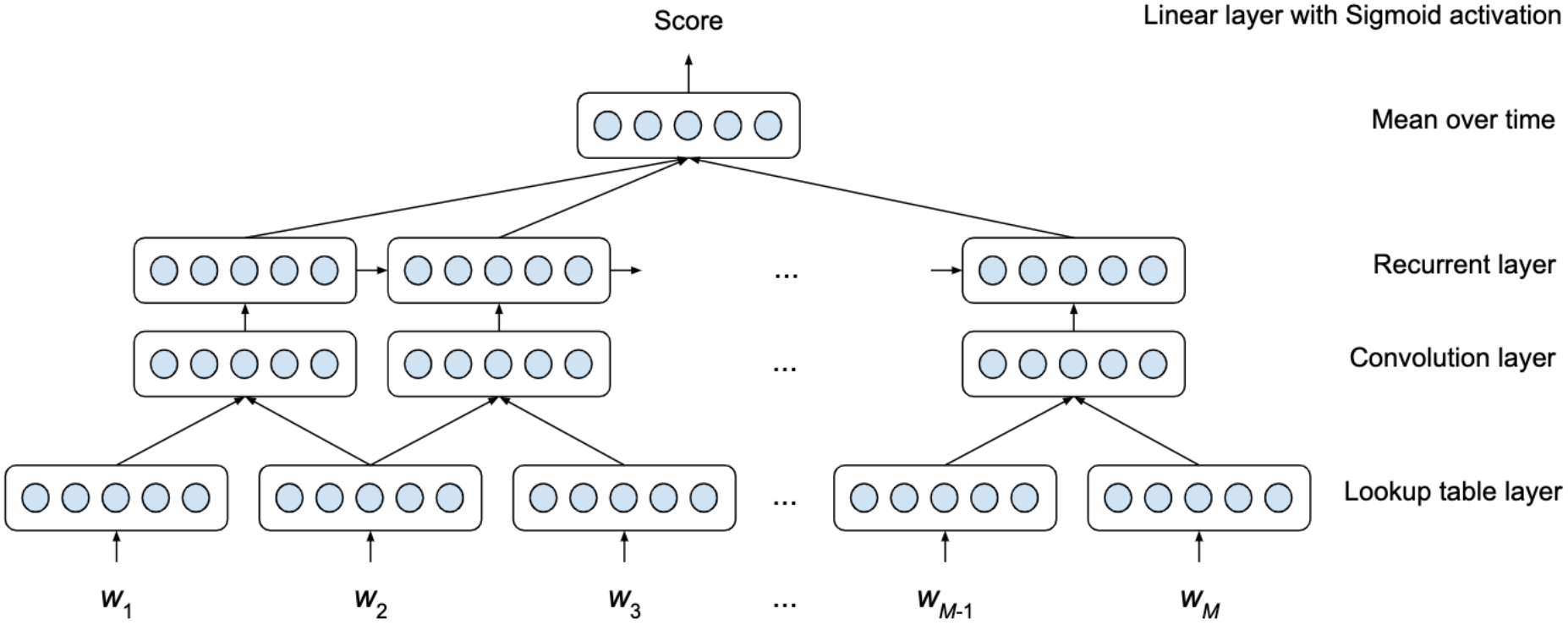
Prompt	Essay Size	Genre	Avg. Len.
1	1783	ARG	350
2	1800	ARG	350
3	1726	RTA	150
4	1772	RTA	150
5	1805	RTA	150
6	1800	RTA	150
7	1569	NAR	250
8	723	NAR	650

Essay instruction prompt 1: More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. Some experts are concerned that people are spending too much time on their computers. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people.

Essay instruction prompt 2: Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive?

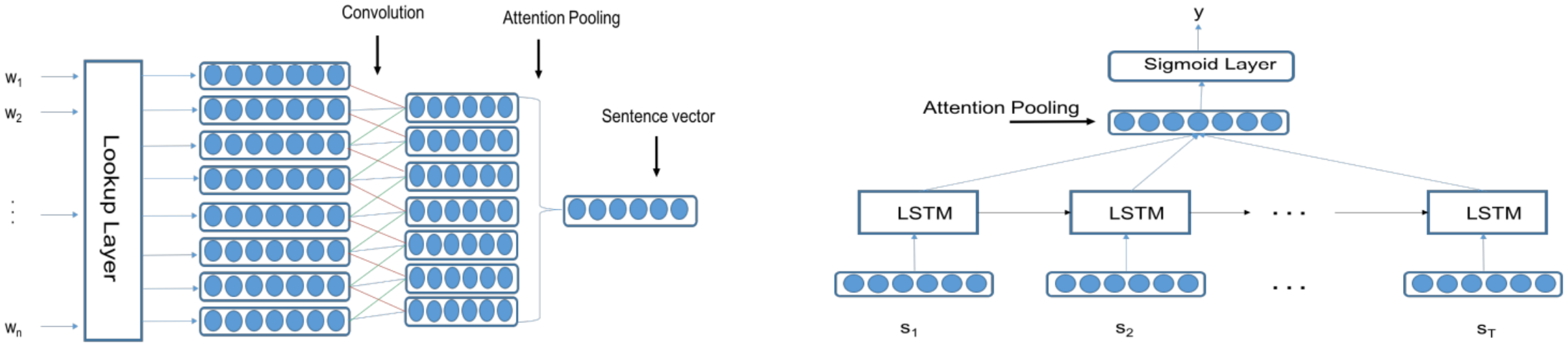
Essay instruction prompt 7: A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.

Neural-based Methods



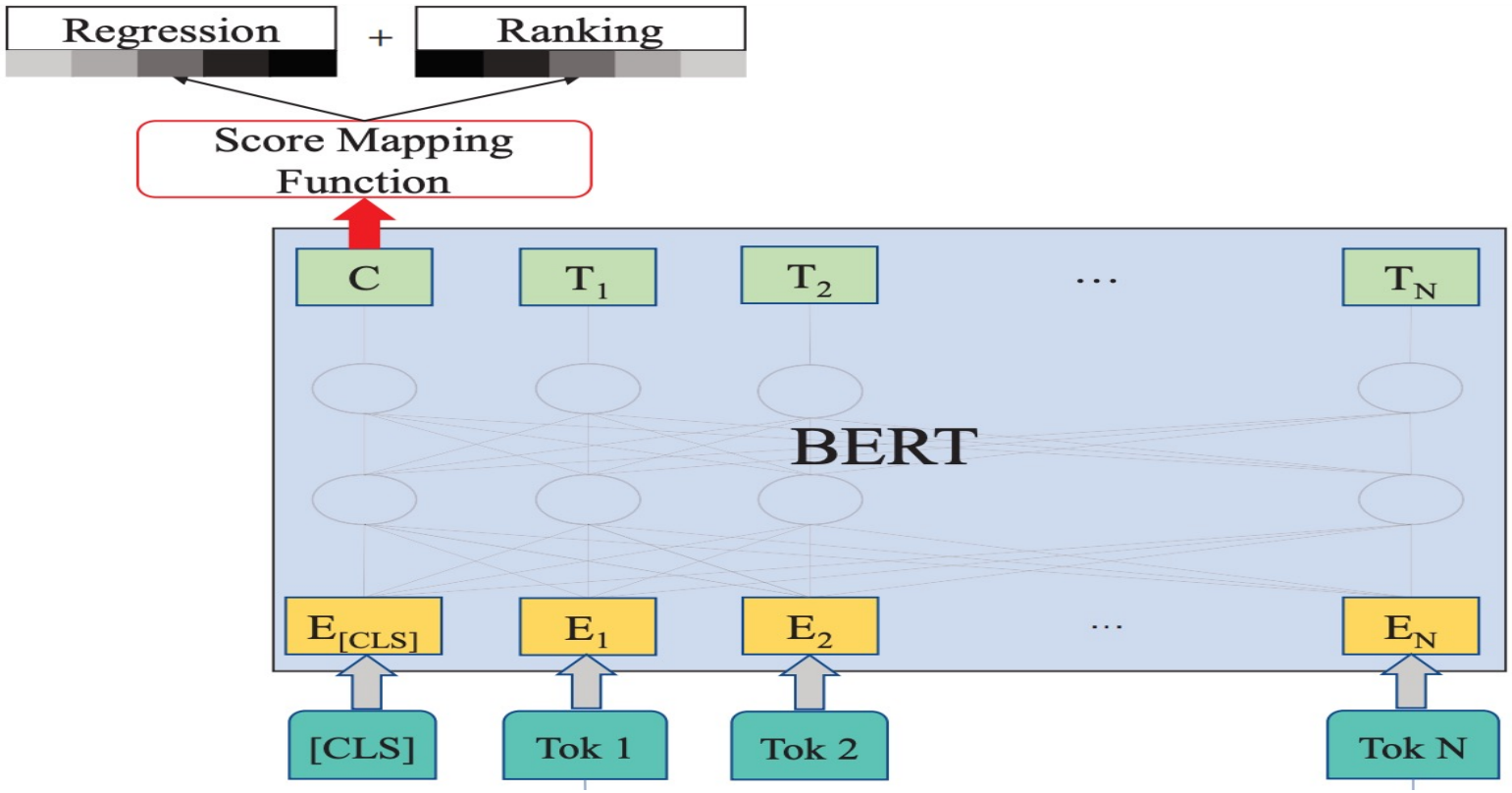
Taghipour et al. 2016 CNN-LSTM based AES

Neural-based Methods



Dong et al. 2017 CNN-LSTM-atten based AES

Neural-based AES systems



Song et al. 2020 BERT-based AES

Current Automated Essay Scoring

Student essays 

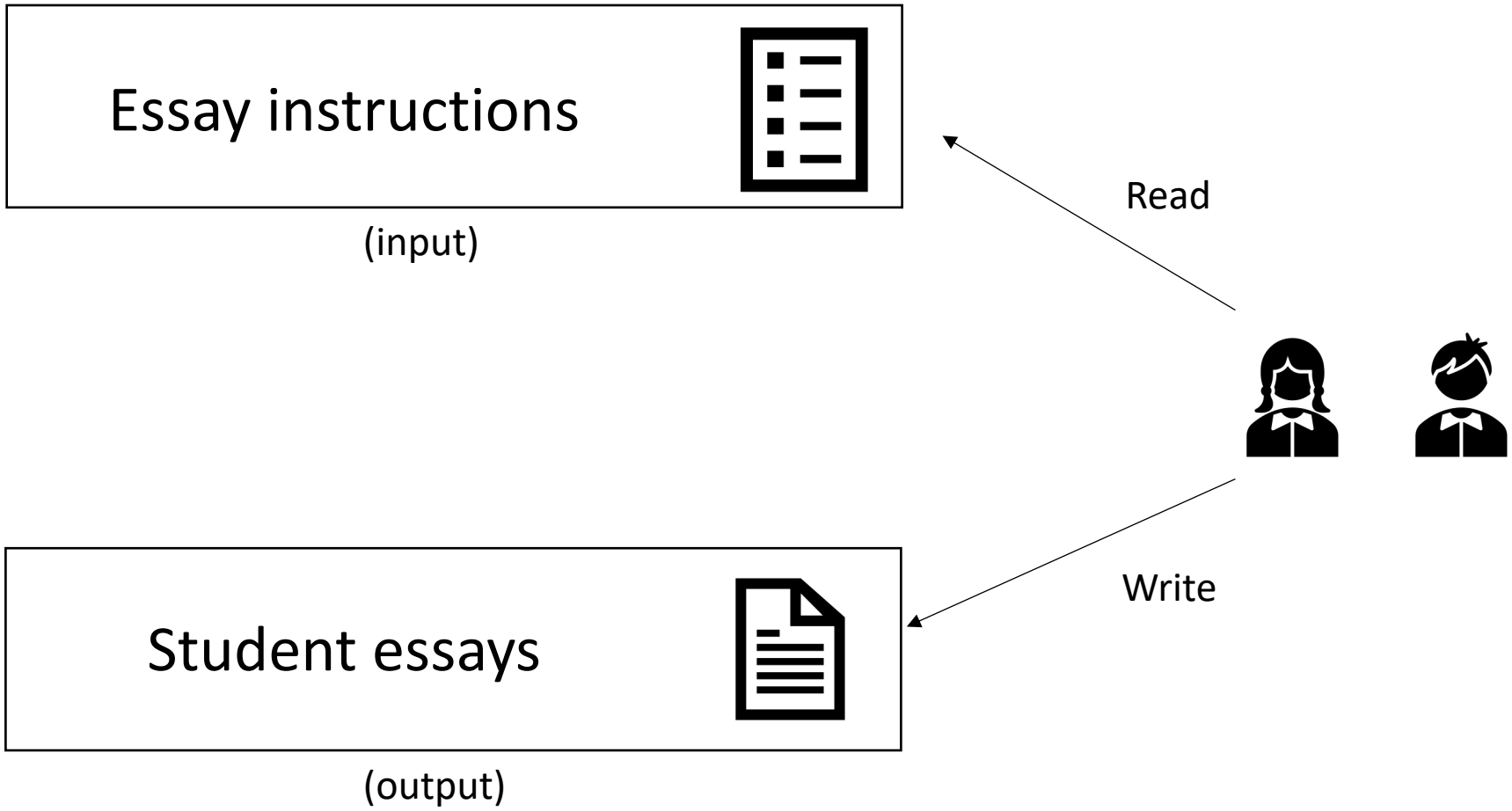
AES System

Holistic Essays Scores

Ruthless
Tireless
Automatic
Consistent

Could we achieve something beyond a holistic predicted score with AES system?

When students write essays



When teachers grade essays

Grading criteria

Student essays

(input)

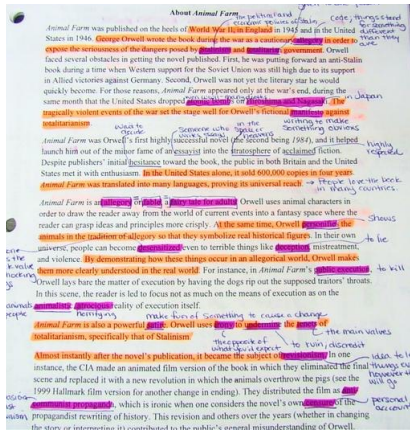
Read

Grade

1. Writing evaluation information
   - Highlighted topical sentences
   - Grammar, etc.

2. Grade

(output)

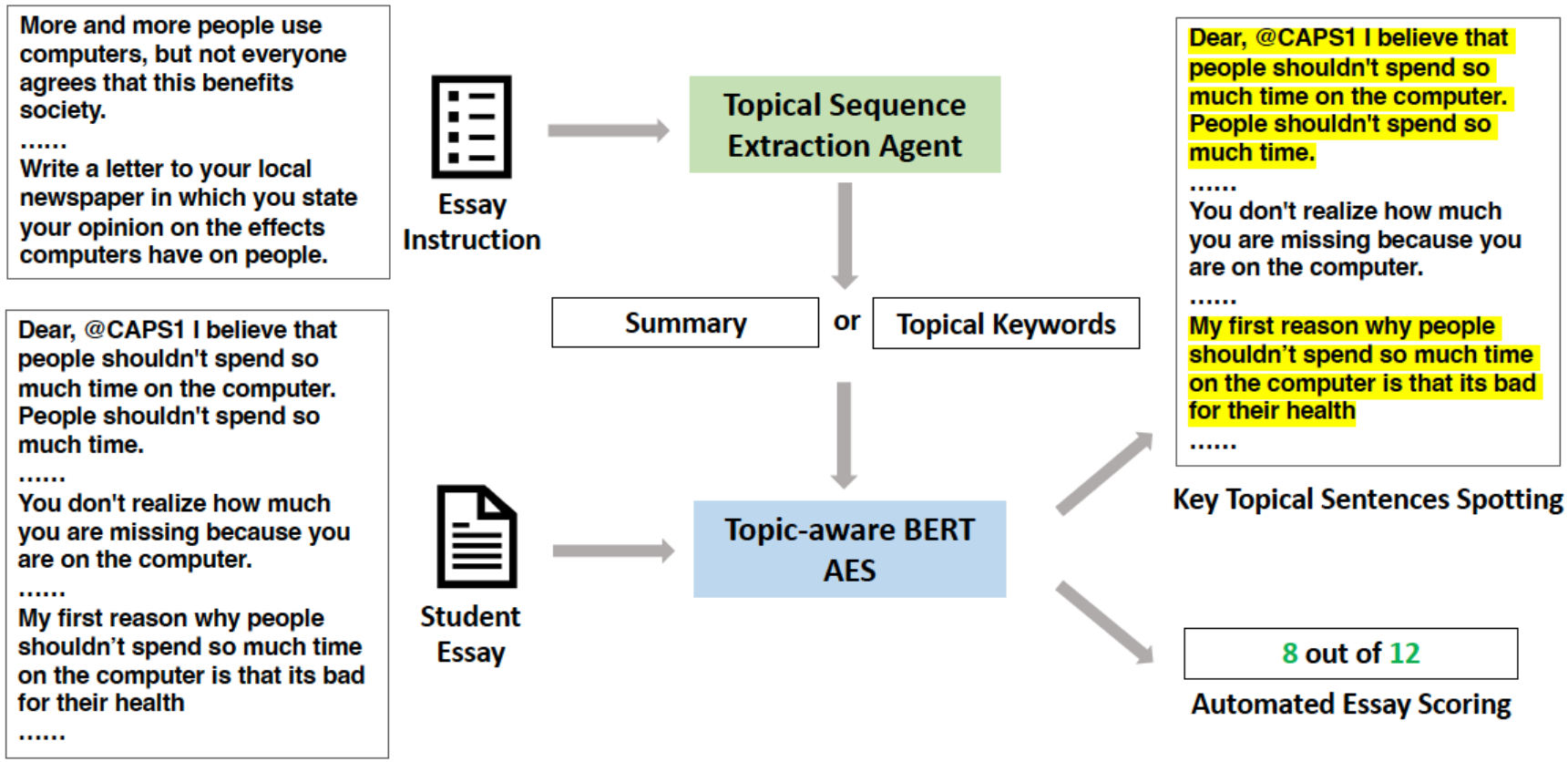


Stockholm University

Beyond Benchmarks: Spotting Key Topical Sentences While Improving Automated Essay Scoring Performance with Topic-Aware BERT

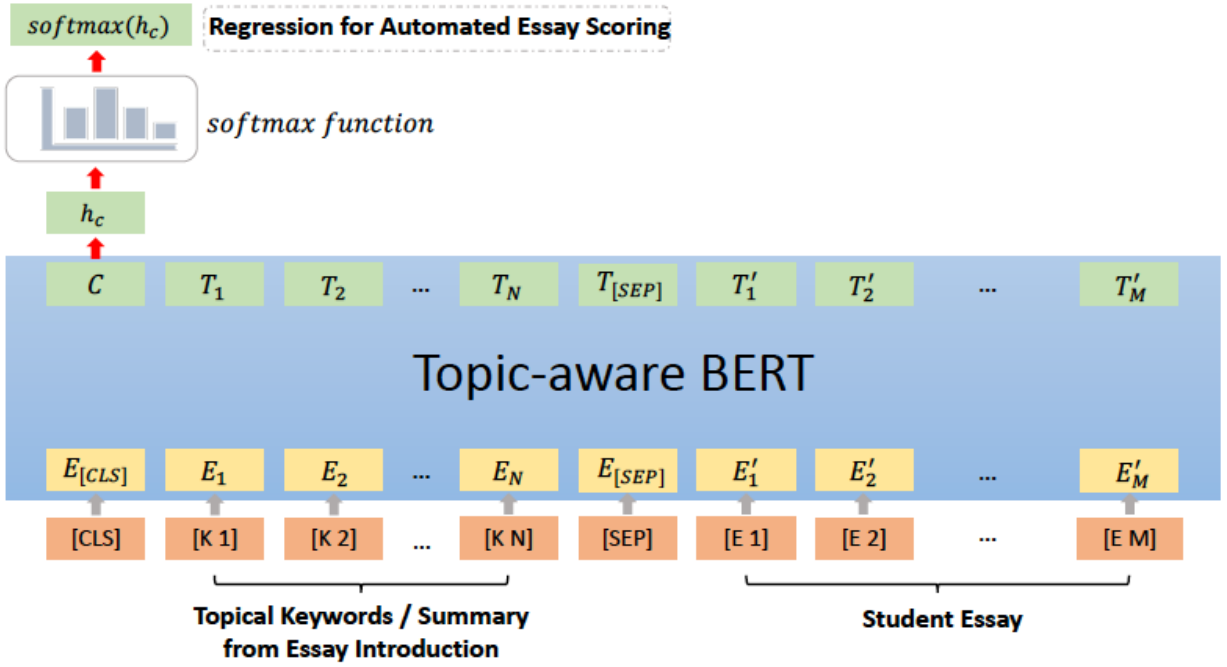
Wu, Y.; Henriksson, A.; Nouri, J.; Duneld, M.; Li, X. 2023

Topic-aware BERT AES system



Topic-aware BERT

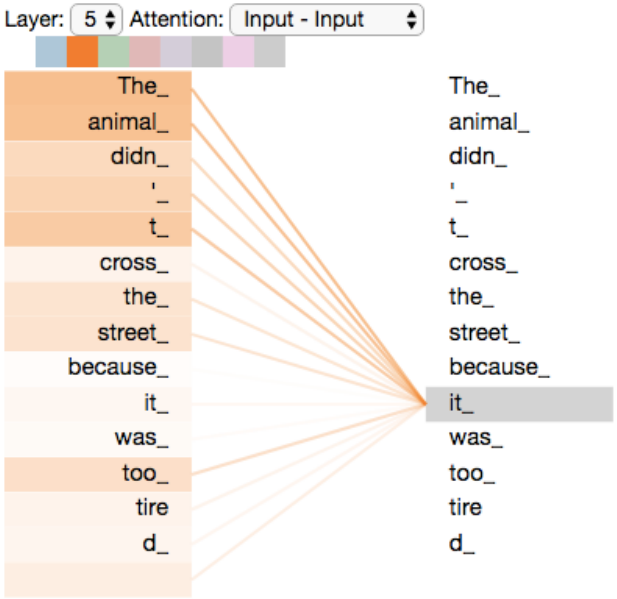
Topical Sequence Extraction Methods	Agent Type	Topical Sequence Type	Topical Sequence Length	Essay Truncation?
Manual	Manual	Topical keywords	4	No
YAKE	Automatic	Key phrases, keywords	32	Yes
Xsum	Automatic	Single-sentence summary	12	No
CNN	Automatic	Multiple-sentence summary	42	Yes

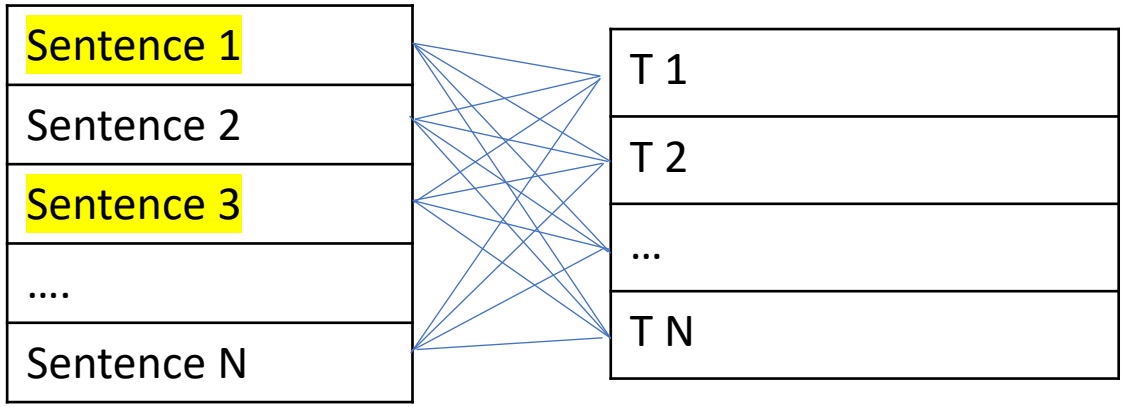


Using self-attention to retrieve KTS

- ”The animal didn't cross the street because it was too tired”

Example from https://jalammar.github.io/illustrated-transformer/





$$Atten\_Sent(s,l) = \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} \alpha_l(t_{s_i}, t_{k_j})}{n}$$

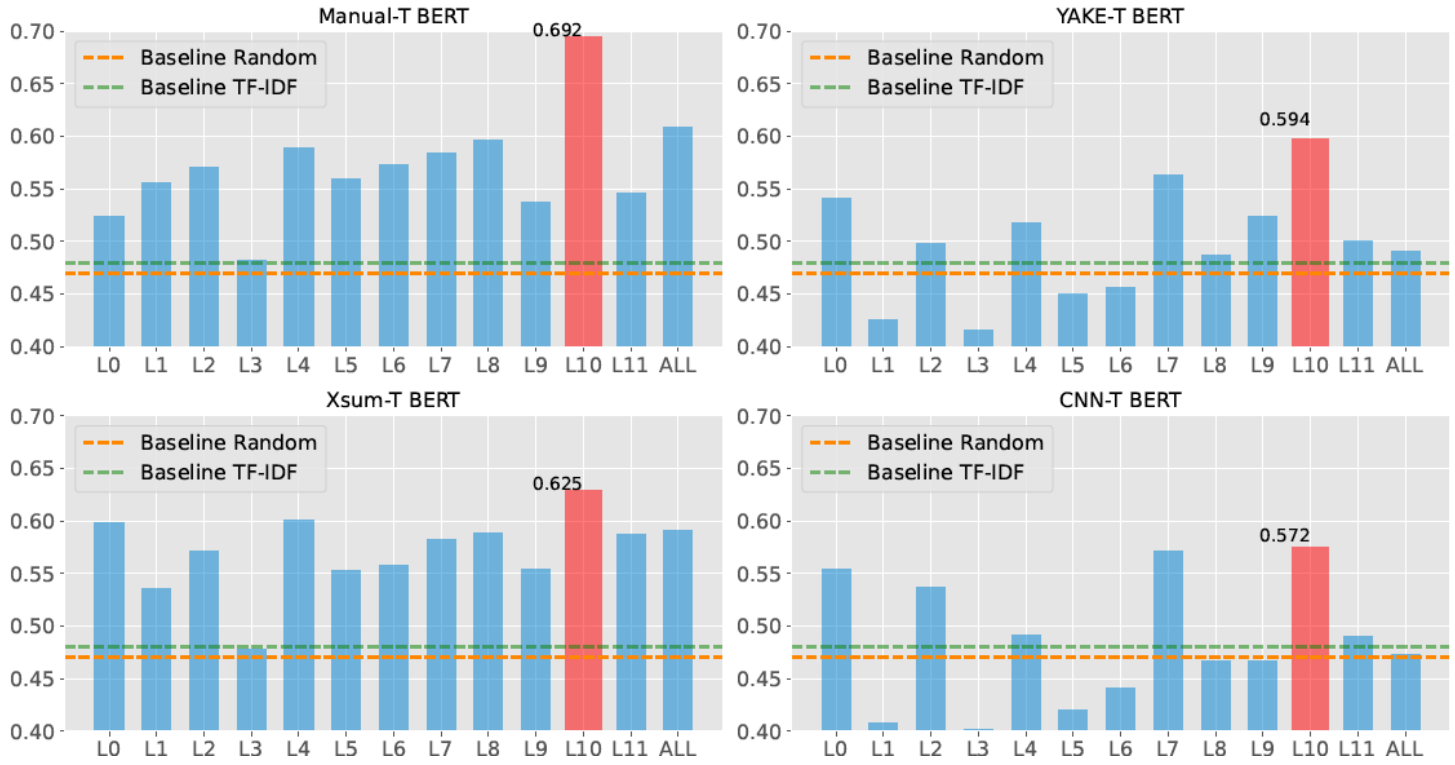
AES performance

Stockholm University

Table 4. QWK scores of different models with essays from prompts one, two, and seven in our experiment. The numbers in the parenthesis indicate the QWK scores reported in the original papers of selected models experimenting with all prompt essays. For example, concerning essays from prompt one, CNN-LSTM achieved 0.789 in our experiments, while the reported QWK in the original paper was 0.821.

	Models	Essay Prompt ID			Average QWK
		1	2	7	
Baselines	CNN-LSTM	0.789 (0.821)	0.687 (0.688)	0.805 (0.808)	0.760 (0.772)
	CNN-LSTM-Att	**0.825** (0.822)	0.658 (0.682)	0.788 (0.801)	0.757 (0.768)
	Vanilla BERT	0.814 (0.821)	0.689 (0.678)	0.820 (0.802)	0.774 (0.767)
Topic-aware BERT	Manual-T BERT	0.822	0.702	0.818	0.781
	YAKE-T BERT	0.813	**0.717**	**0.837**	**0.789**
	Xsum-T BERT	0.821	0.710	0.836	**0.789**
	CNN-T BERT	0.803	0.714	0.833	0.783

KTS retrieving performance



$$MAP = \frac{1}{Q}\sum_{j=1}^{Q}\frac{1}{m_j}\sum_{k=1}^{n} P@k \times rel@k$$

$$P@k = \frac{\{\text{key topical sentences}\} \cap \{\text{retrieved sentences}\}}{\{\text{retrieved sentences}\}}$$

The future

Longformer?

Detect essay generated from ChatGPT?

Stockholm University

QUESTIONS ?