Effectiveness of Data Augmentation and Pretraining for Improving Neural Headline Generation in Low-Resource Settings

Matej Martinc[1], Syrielle Montariol[2], Lidia Pivovarova[3], Elaine Zosa[3]

[1]Jozef Stefan Institute, [2]INRIA Paris, [3]University of Helsinki

Headline Generation

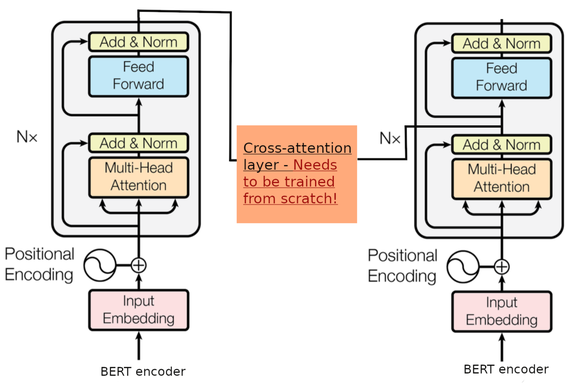
- Use Transformer Encoder-Decoder architectures, popular for **machine translation** and **summarization**.
- Headlines are a vehicle that carries the most important information about the event or topic described in the news article ⇒ **Headline generation as a summary**

Two SOTA summarization systems

BERT-ED

BERT encoder-decoder

- Combination of any pretrained BERT language models.
- **Can be employed for 168 languages.**



BART

Transformer-based model, pre-trained on denoising tasks, sentence shuffling and text infilling.

- **Multilingual version supports 50 languages.**
- **Potential problem: curse of multilinguality?**

Pre-training

- **Sentence shuffling**: Randomly shuffle sentences, train the model to generate the original text with the correct sentence order.
- **Text infilling**: Corrupt 20% of the training corpus by an infilling scheme (spans of text are replaced with a [MASK] token), then train the encoder-decoder to generate the original text.
- **2 tasks**: Train the model with Sentence shuffling then Text infilling.

Data Augmentation

- **BERT**: Mask 20% of words. Feed the masked article to BERT, which proposes candidate words for the masked tokens.
- **Word2Vec**: Replace random words by synonyms proposed by Word2Vec.
- **Wordnet**: Same as above but using Wordnet.
- **EDA**: Perform synonym replacement, random insertion, random swap, and random deletion.
- **Mixed**: Successively perform Word2Vec, EDA and Wordnet augmentation.

For each article in the train set, generate 5 augmented articles.

Evaluation metrics

- **ROUGE 1, 2, L**: Current standard for evaluating summarization. Agnostic to semantic similarity, hence can have low correlation with human judges.
- **Semantic similarity (SS)**: measures cosine distance between the sentence embedding of the true and generated headline.
- **Natural language inference (NLI)**: consider the true headline as the 'premise' and the generated headline as the 'hypothesis'; use an NLI model to predict entailment, as a measure of generated headline quality. Only used for English experiments!

Results

Approach		ROUGE-1		ROUGE-2		ROUGE-L		SS		NLI	
English BERT-ED-based models											
BASELINE	10K	10.2		1.4		9.6		24.6		15.4	
	260K	27.6		10.1		25.1		49.6		32.1	
AUGMENTATION	bert	13.2	3.0	2.3	0.9	12.2	2.6	30.8	6.2	15.7	0.3
	w2v	9.7	-0.5	1.6	0.2	8.9	-0.7	26.5	1.9	14.9	-0.5
	mix	10.4	0.2	1.7	0.3	9.6	0.0	23.9	-0.7	13.1	-2.3
PRETRAINING	2 tasks	**16.5**	**6.3**	**4.6**	**3.2**	**15.1**	**5.5**	**42.0**	**17.4**	**25.9**	**10.5**
English BART-based models											
BASELINE	10K	**29.0**		**10.9**		**26.0**		49.3		34.1	
	260K	31.9		13.1		28.7		**51.7**		**36.8**	
AUGMENTATION	bert	28.5	-0.5	10.5	-0.4	25.6	-0.4	49.1	-0.2	34.0	-0.1
	w2v	27.8	-1.2	10.1	-0.8	25.1	-0.9	48.2	-1.1	32.0	-2.1
	mix	27.7	-1.3	10.2	-0.7	25.0	-1.0	47.9	-1.4	32.2	-1.9
PRETRAINING	2 tasks	28.7	-0.3	10.7	-0.2	25.9	-0.1	49.2	-0.1	34.1	0.0

Approach		ROUGE-1		ROUGE-2		ROUGE-L		SS	
Estonian BERT-ED-based models									
BASELINE		3.9		0.3		3.8		17.9	
AUGMENTATION	bert	9.8	5.9	2.5	2.2	9.4	5.6	36.9	19.0
	w2v	8.5	4.6	2.1	1.8	8.1	4.3	34.4	16.5
PRETRAINING	infilling	13.9	0.1	4.3	4.0	13.2	9.4	44.0	26.1
	shuffling	11.3	7.4	2.8	2.5	10.7	6.9	40.7	22.8
	2 tasks	**17.6**	**13.7**	**6.5**	**6.2**	**16.3**	**12.5**	**49.8**	**31.9**
Estonian BART-based models									
BASELINE		26.2		12.3		24.4		56.7	
AUGMENTATION	bert	25.4	-0.8	11.6	-0.7	23.8	-0.6	55.9	-0.8
	w2v	23.0	-3.2	9.8	-2.5	21.5	-2.9	53.5	-3.2
PRETRAINING	infilling	**27.1**	**0.9**	**12.9**	**0.6**	**25.2**	**0.8**	**57.2**	**0.5**
	shuffling	26.6	0.4	12.6	0.3	24.8	0.4	56.9	0.2
	2 tasks	26.6	0.4	12.3	0.0	24.6	0.2	56.6	-0.1

Table: Results on the English and Estonian datasets. Best results in a low resource setting (i.e., excluding the BART and BERT-ED models trained on English 260K dataset) per evaluation measure and language are **bolded**. For each measure, we report its absolute value (1st number) and the difference with the baseline model (2nd, colored number). Since all experiments with data augmentation and pretraining are run on the 10K dataset, differences are computed respectively to the 10K baseline, the first row for each model.

Examples

	Example
True headline	fighting n. y. c. soda ban, industry focuses on personal choice
BART 260K	soda industry fights new york city's soda ban
BART 10K	soft-drink industry takes aim at sugary drinks
BERT-ED 260K	soft - drink industry seeks to fight sugary drinks ban on sugary drinks
BERT-ED 10K	in new yorks's york taxs's taxs s
BERT-ED 10K + BERT aug	in new york city, new york city's new york city's bans law
BERT-ED 10K + shuffling	u. s. and new york's new new york city mayor' campaign campaign moves new york's mayor's campaign campaign
BERT-ED 10k + infilling	philippines : s. o. p. to be suspended s. a. lawmakers s. ban s. a.'s
BERT-ED 10K + 2 tasks	new yorkers face a challenge to soda industry in new yorkers in new yorkers' campaign campaign in new york city's

Table: Examples of generated English headlines.

- BART 10k produces results very similar to BART 260k but can hallucinate.
- Performance of BERT-ED-based model trained on 10K drops compared to the 260K dataset.
- Data augmentation only slightly improves the performance on English.
- Pretraining on 2 tasks works the best, resulting in longer headlines with meaningful beginnings.

Datasets

News datasets for evaluation of headline generation:

Language	train set	test set
English KPTimes 260K - BASELINE	259,923	10,000
English KPTimes 10K	10,000	10,000
Croatian	32,223	3,582
Estonian	10,750	7,747

Conclusion

- Pretrained multilingual NLG model for a specific low-resource language should be picked over training the cross-attention layer from scratch.
- Pretraining and data augmentation has little effect on BART-based models.
- Pretraining > augmentation.
- SS and NLI are highly correlated w/ ROUGE.
- The paper repository:
- More details: MATEJ.MARTINC@IJS.SI