Importance of Copying Mechanism for News Headline Generation

Ilya Gusev, PhD student

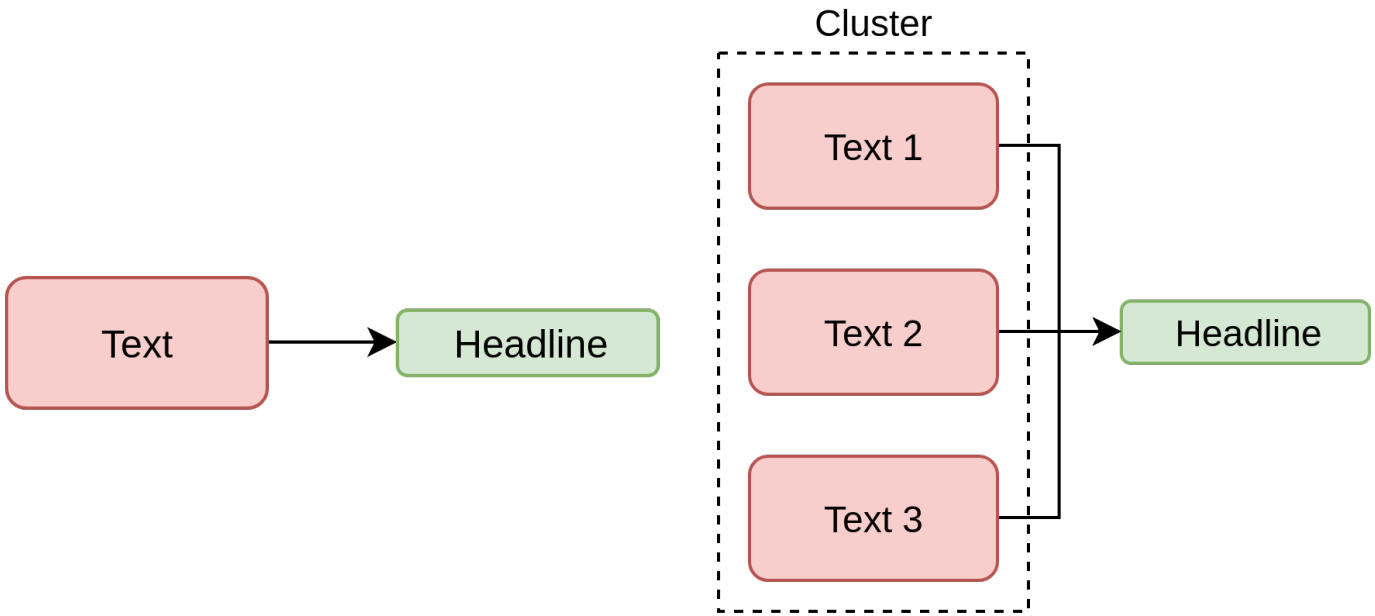
Moscow Institute of Physics and Technology
Department of computational linguistics

Moscow, 30 May 2019

Ultimate task of generating headline for news

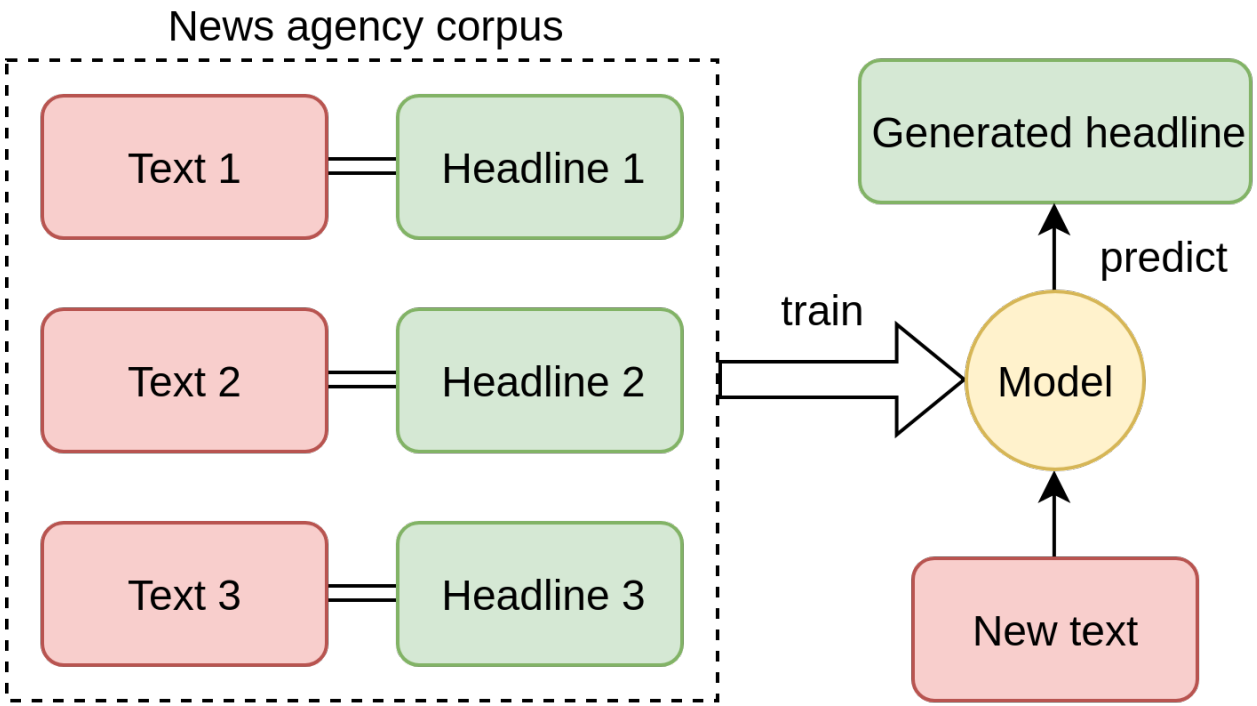
Given a text or a collection of texts from news produce a single sentence that is:

- grammatically correct
- capturing main information
- not too short and not too long
- interesting for readers
- unbiased



Current research task

Predict a headline for a new text document of a particular news agency given a collection of texts with headlines of this agency.



Datasets

Russian news agencies:

1. RIA dataset:
   - 1 million documents
   - from 2010 to 2014.
2. Lenta dataset:
   - 800 thousand documents
   - from 1999 to 2018
3. ROMIP dataset:
   - 32 thousand documents

What is wrong with datasets?

1. No timestamps in RIA and Lenta documents $\Rightarrow$ no time based splits $\Rightarrow$ entity bias
2. HTML markup in RIA documents $\Rightarrow$ final scores depend on quality of preprocessing
3. Only one headline $\Rightarrow$ agency bias

Baseline

Baseline: take a first sentence of the input text as a headline

	R-mean-f	BLEU
RIA	17.12	21.22
Lenta	18.59	25.45

Table: Baseline scores on test sets

Seq2seq with attention

First introduced by Bahdanau et al. in 2014 for machine translation.

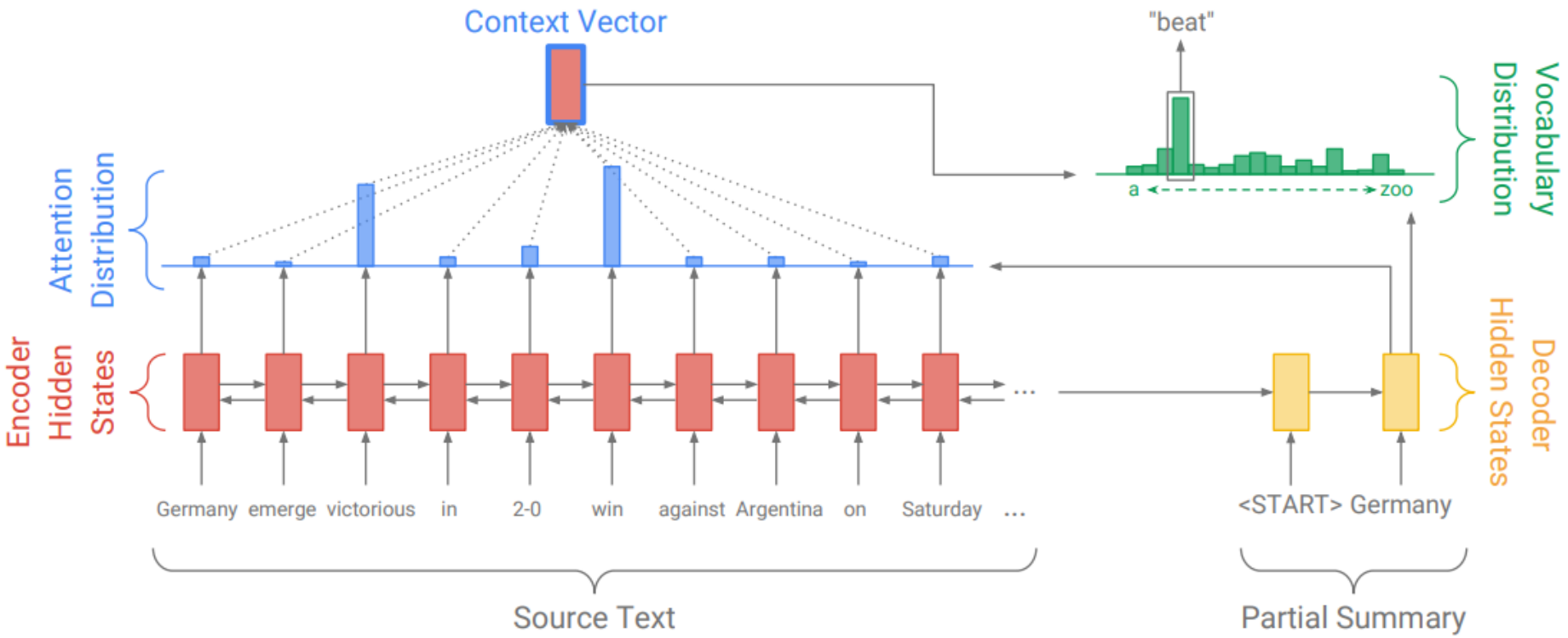


Figure: Seq2seq model with attention. Illustration from See et al., 2017

CopyNet

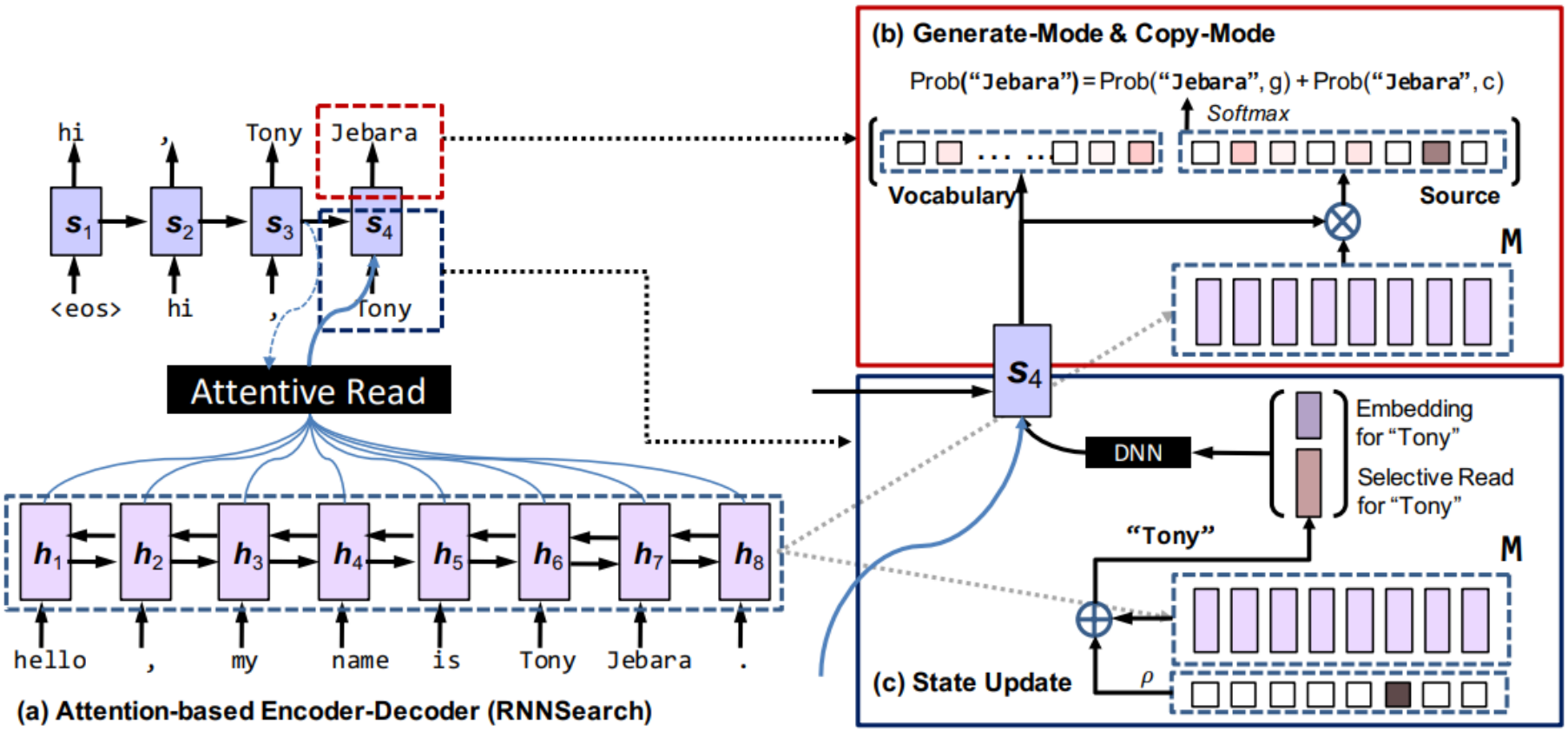


Figure: CopyNet. Gu et al., 2016

Pointer-Generator Networks

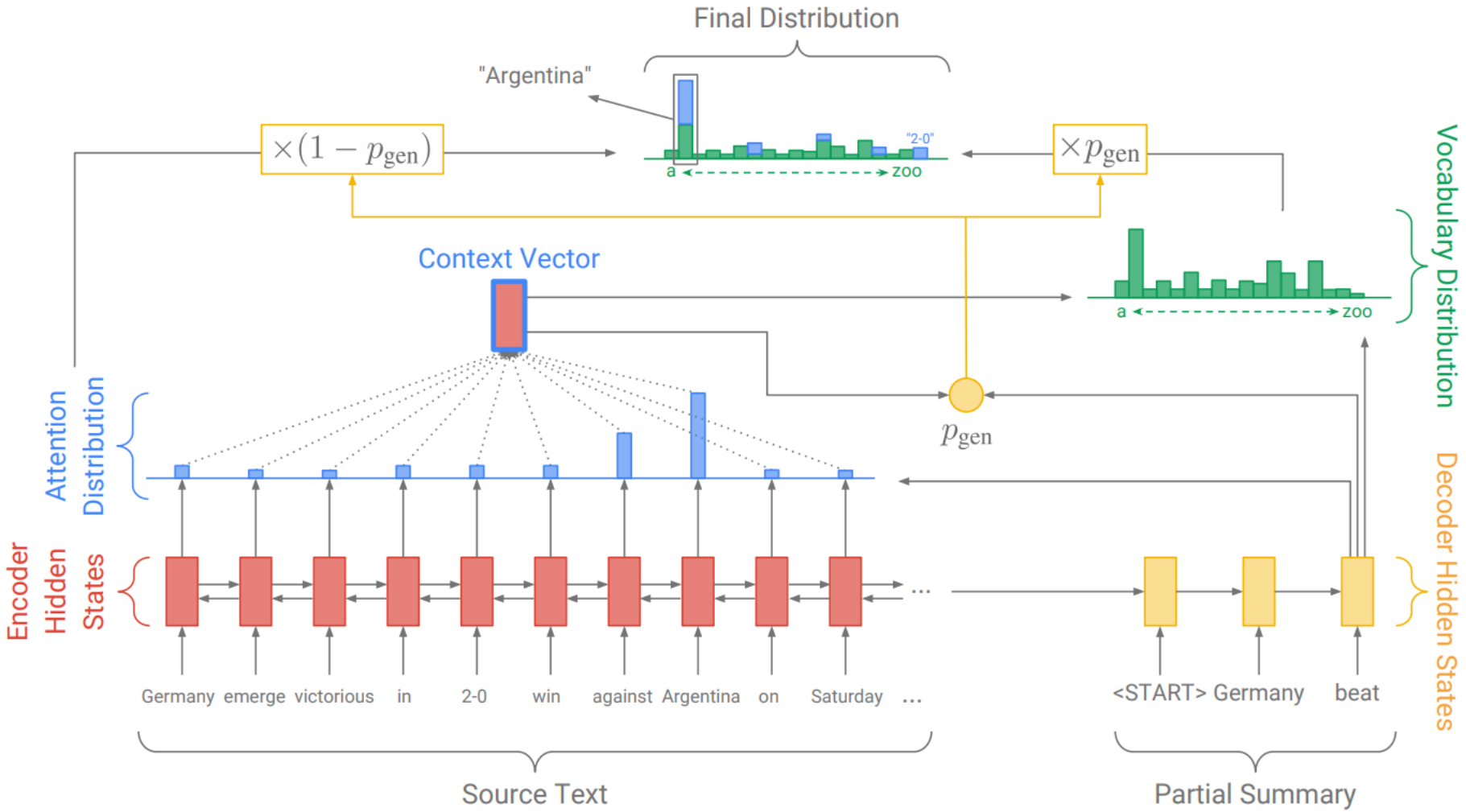


Figure: Pointer-Generator Networks. See et al., 2017

Byte pair encoding tokenization

Iteratively take most frequent bigram as new token. Useful for morphologically-rich languages, like Russian, as it encodes frequent suffixes and prefixes as tokens.

1. aaabdaaabac
2. ZabdZabac, Z=aa
3. ZYdZYac, Y=ab, Z=aa
4. XdXac, X=ZY, Y=ab, Z=aa

Technical details

1. Python 3.6, PyTorch 1.0.0, AllenNLP 0.8.2, sentencepiece 0.1.8
2. Custom dataset readers for RIA, Lenta and CNN/DailyMail
3. LSTM as encoder and decoder
4. PGN: self-written
5. CopyNet: AllenNLP
6. BPE: sentencepiece

RIA results

Model	R-mean-f	BLEU
seq2seq-bpe-5m	32.20	49.77
copynet-words-10m	33.00	51.99
copynet-bpe-10m	33.57	52.57
seq2seq-words-25m	30.31	44.69
seq2seq-bpe-25m	33.58	51.66
copynet-words-25m	33.81	52.99
pgn-words-25m	33.64	51.48
pgn-subwords-24m	33.73	52.61
copynet-bpe-43m	34.97	53.80
First Sentence	17.12	21.22
UTransformer (Gavrilov et al., 2019)	32.90	-
PBA Transformer (Sokolov, 2019)	36.14	-

Table: RIA dataset evaluation

RIA train, Lenta and ROMIP test

Model	Lenta R-mean-f	ROMIP R-mean-f
seq2seq-bpe-5m	14.86	14.85
seq2seq-words-25m	13.91	-
seq2seq-bpe-25m	15.89	15.40
pgn-subwords-24m	17.00	-
pgn-words-25m	17.15	-
copynet-words-10m	21.02	-
copynet-words-25m	22.53	-
copynet-bpe-10m	20.32	21.69
copynet-bpe-43m	**22.66**	**23.00**
First sentence	18.59	19.50

Table: Lenta and ROMIP datasets evaluation with a model trained on RIA dataset

Example

Reference title	дело в отношении бельтюкова не скажется на "сколково" - вексельберг
seq2seq-words-25m	"сколково" UNK возбуждение дела против UNK
seq2seq-bpe-5m	"сколково" обеспокоен возбуждением дела против экс-главы фонда
seq2seq-bpe-25m	"сколково" считает возбуждение дела против вице-президента
copynet-words-10m	ситуация против бельтюкова не скажется на воплощении проекта "сколково"
copynet-bpe-10m	"сколково" озабочено возбуждение дела против бельтюкова
copynet-bpe-43m	руководство "сколково" озабочено возбуждение дела против бельтюкова

Table: Beltukov example

Example

Reference title	лучшей смерти он себе и не желал – вдова журналиста марка дейча
seq2seq-words-25m	UNK
seq2seq-bpe-5m	день на бали
seq2seq-bpe-25m	марк дейча
copynet-words-10m	"за заслуги перед отечеством": "за заслуги перед отечеством"
copynet-bpe-10m	друзья и родные проводили в последний путь журналиста марка дейча
copynet-bpe-43m	марк дейч утонул на острове бали

Table: Mark Deutch example

Conclusions

- The copying mechanism is required for the headline generation task
- The model trained on news of one agency generate decent headlines for the texts of another agency
- There are still errors in generated headlines

Better experiment designs

Algorithm:

1. Parse news from many agencies
2. Produce records with text-title-agency-timestamp structure
3. Perform text-based clustering within fixed time frames
4. Split dataset using timestamps
5. First way:
   - Learn to generate a title based on a corresponding text
   - Use all titles during evaluation
6. Second way:
   - Learn to generate every title from every text in the cluster
   - Use agency as one of the inputs
7. Third way:
   - Learn to generate every title using all texts from the cluster simultaneously
8. Forth way:
   - Rank titles
   - Learn to generate the best title using all texts from the cluster simultaneously

Useful links I

- **Article**
  Importance of Copying Mechanism for News Headline Generation
- **Project repo**
  https://github.com/IlyaGusev/summarus
- **E-mail**
  phoenixilya@gmail.com
- **Telegram**
  YallenGusev