Automatic Generation of Factual News Headlines in Finnish

Maximilain Koppatz, Khalid Alnajjar, Mika Hämäläinen, Thierry Poibeau

Finnish NLG

- Finnish NLG research has relied on rules
- Lack of pretrained neural models for Finnish (during our research)

News headlines

- News outlets need to write headlines all the time
  - Multiple candidates for each news article
- A/B testing

Collaboration with Sanoma

- One of the largest media houses in Finland
- Daily news paper
  - Helsingin Sanomat
- Daily yellow press paper
  - Ilta-Sanomat

Data

- Sanoma corpus
  - around 3.8 million Sanoma news articles published between the year 1990 and 2021
  - headline, ingress, main text
  - used for headline generation
- Other corpora for pretraining
  - The Finnish Wikipedia, around 460,000 pages
  - Yle news corpus, around 100,000 articles
  - Ylilauta corpus, around 335,000 messages

The GPT-2 Model

- Tokenizer
    - Byte-pair-encodings
    - Vocabulary size 50,000
- GPT-2
    - The entire corpus (Sanoma, Yle, Ylilauta, Wikipedia) without headlines

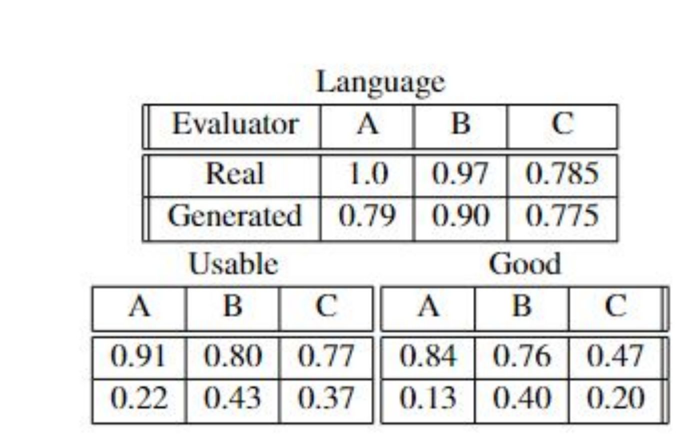
Fine-tuning for headline generation

- Headline generation as summarization
- Body of the news article (max 448 tokens) + <special1> + headline + <eos>
  - Model's maximum 512 tokens
- Diverse beam search
  - Gaussian process optimization for hyperparameters with BLEU as objective function

Human evaluation

- 100 news articles from Helsingin Sanomat, 100 from Ilta-Sanomat
- 4 generated headlines for each news story + 1 human written headline (random order)
- Pass-fail
  - Language
  - Usable
  - Good (publication ready)
- Two editors from Ilta-Sanomat, 1 from Helsingin Sanomat

Results



Language

Evaluator	A	B	C
Real	1.0	0.97	0.785
Generated	0.79	0.90	0.775

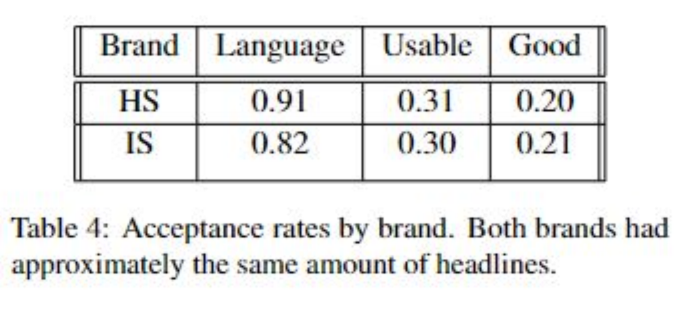
Usable

A	B	C
0.91	0.80	0.77
0.22	0.43	0.37

Good

A	B	C
0.84	0.76	0.47
0.13	0.40	0.20

Results Helsingin Sanomat vs Ilta-Sanomat



Brand	Language	Usable	Good
HS	0.91	0.31	0.20
IS	0.82	0.30	0.21

Table 4: Acceptance rates by brand. Both brands had approximately the same amount of headlines.

Thank you