A zero-resource approach for CS text generation and filtering for automatic speech recognition

Olga Iakovenko (UoS), M. Umar Farooq (UoS), Jie Chi (UoE), Elaf Islam (UoS), Brian Lu (JHU)

Mentor: Hexin Liu

Problem

Little textual CS data available for training satisfactory LMs for CS ASR systems.

Automatic CS text generation does not take into account some of the prior linguistic knowledge.

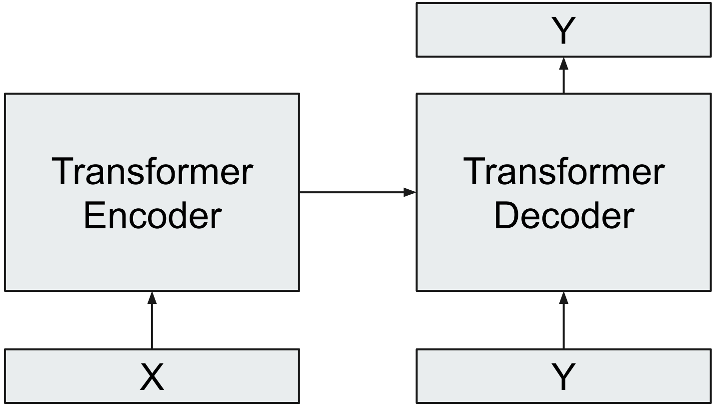
okay kay 让我拿出我的 calculator

Solution

I. Generate arbitrary CS data using constrained beam search decoding within an NMT system trained with parallel data.

II. Automatically select the examples from the data which seem as the most natural CS examples

I. Arbitrary CS text generation

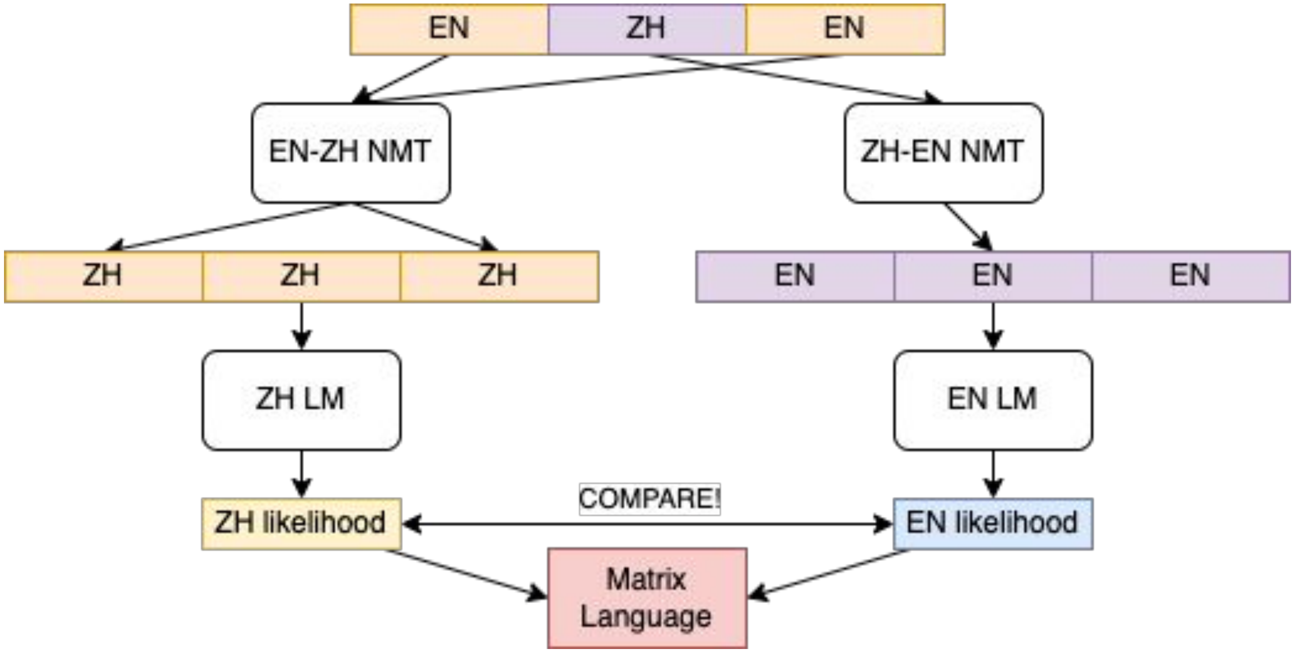


Transformer encoder-decoder translation

Constrained beam search: token masking for creating CS points

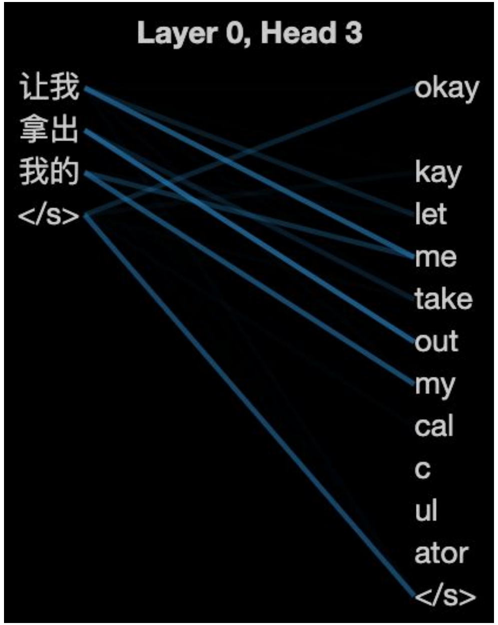
II. Synthetic data filtering

Idea 1: There is a main (Matrix) language that defines the grammar



II. Synthetic data filtering

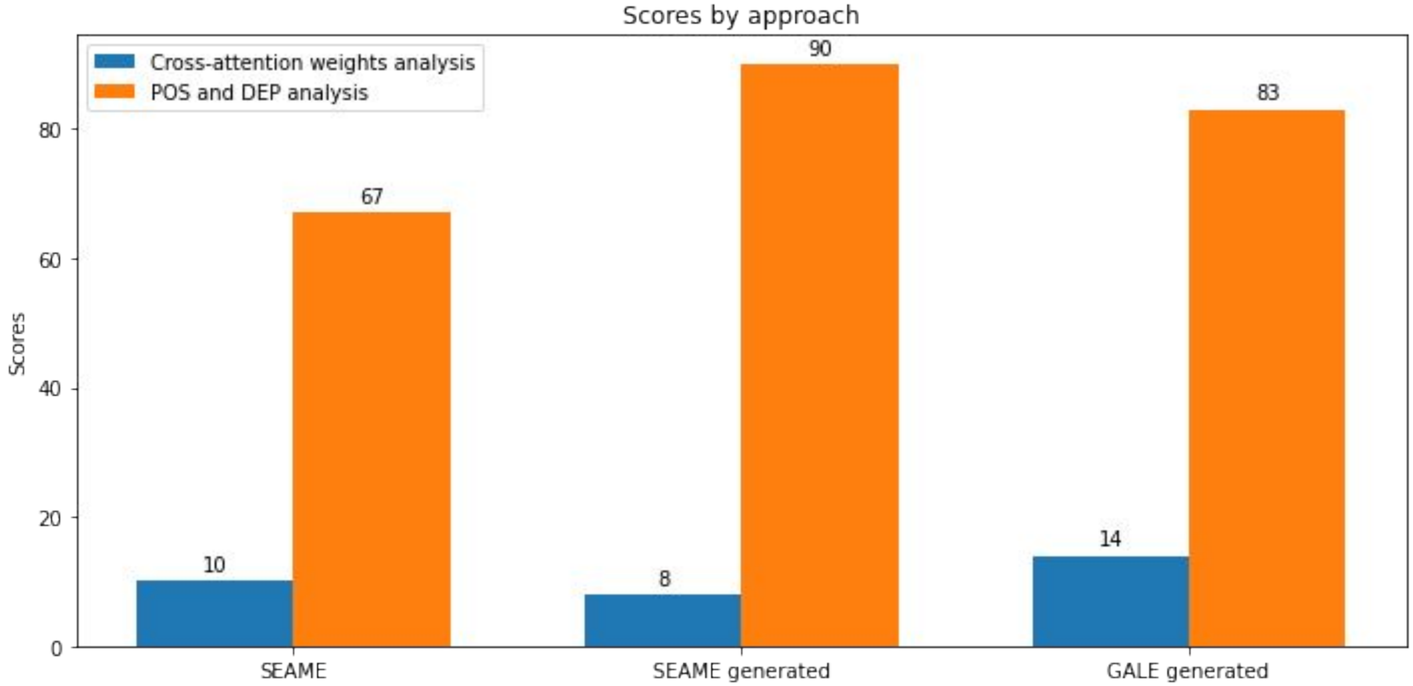
Idea 2: Some words & word categories tend to switch more



Cross-attention analysis of incomplete translations

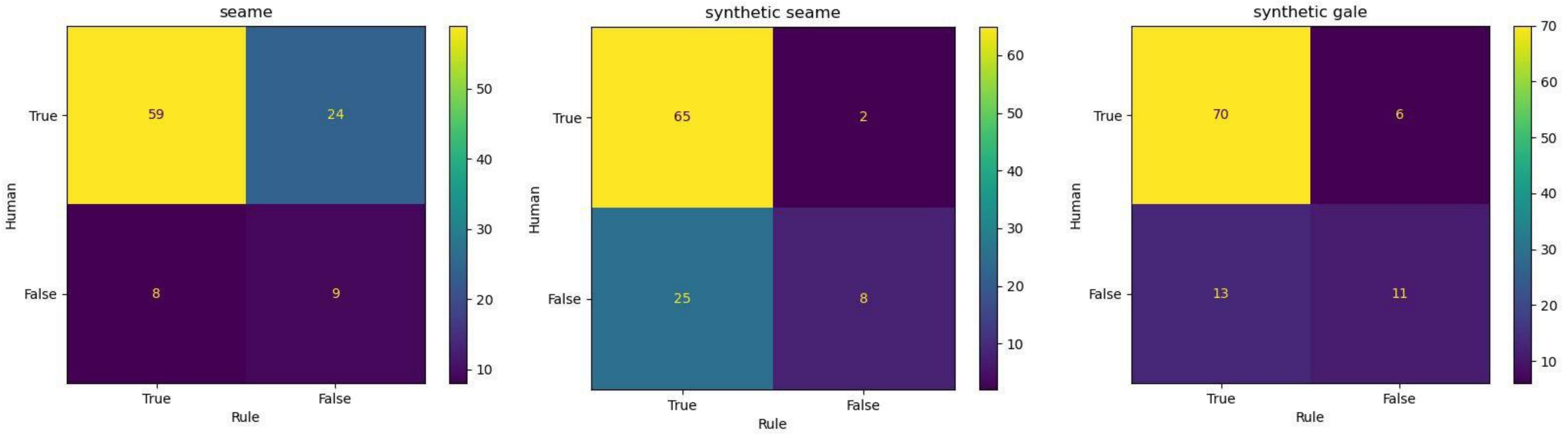
POS tag & dependency analysis: function words do not appear in isolation!

Results



Scores by approach
Cross-attention weights analysis
POS and DEP analysis
Scores
0
20
40
60
80
10
67
8
90
14
83
SEAME
SEAME generated
GALE generated

Results


seame
True
False
Human
59
24
8
9
True
False
Rule
50
40
30
20
10
synthetic seame
True
False
Human
65
2
25
8
True
False
Rule
60
50
40
30
20
10
synthetic gale
True
False
Human
70
6
13
11
True
False
Rule
70
60
50
40
30
20
10

Thank you!

Demo of the project: https://t.me/cs_assessment_bot

@CS_ASSESSMENT_BOT