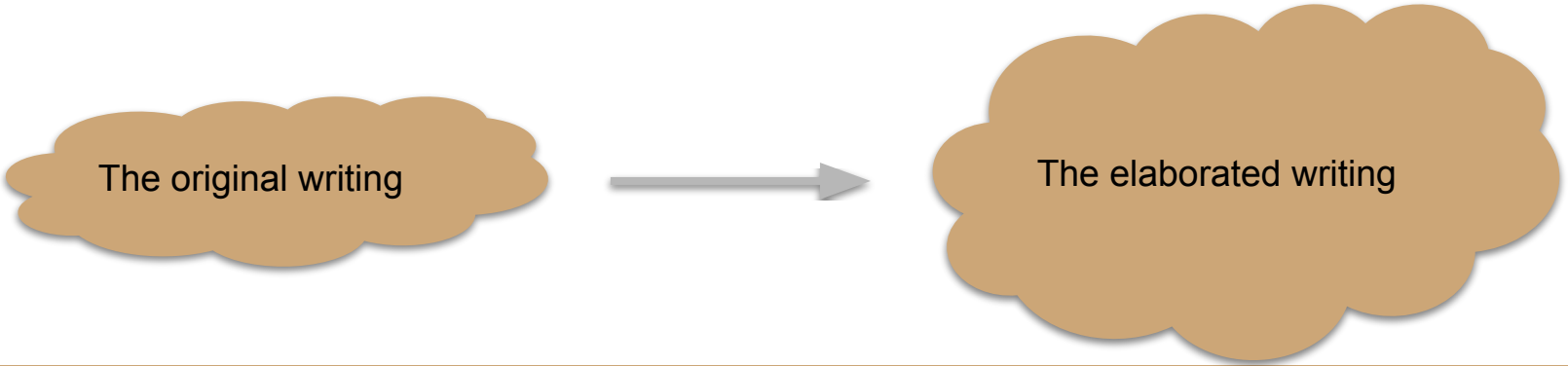
Save Your Words

Mikhail Tokarev
Nuttaree Busarapongpanich
Rashmi Jadhav
Yu-Wen Chen

The Objective

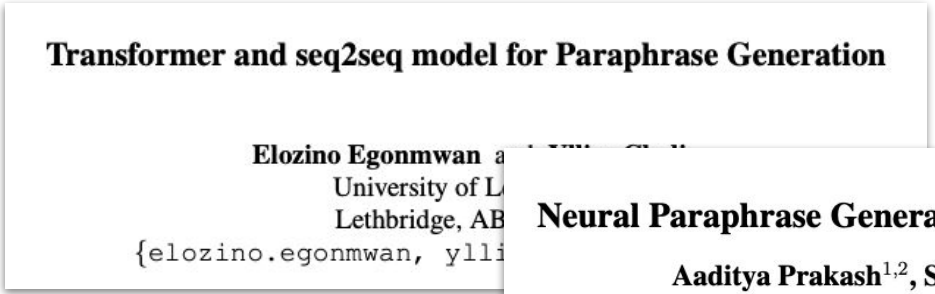
- Paraphrase sentences and make them longer
  - Preserve the original semantics
- Fun experiment to expand sentences
  - detailed and more text is useful for understanding a topic thoroughly



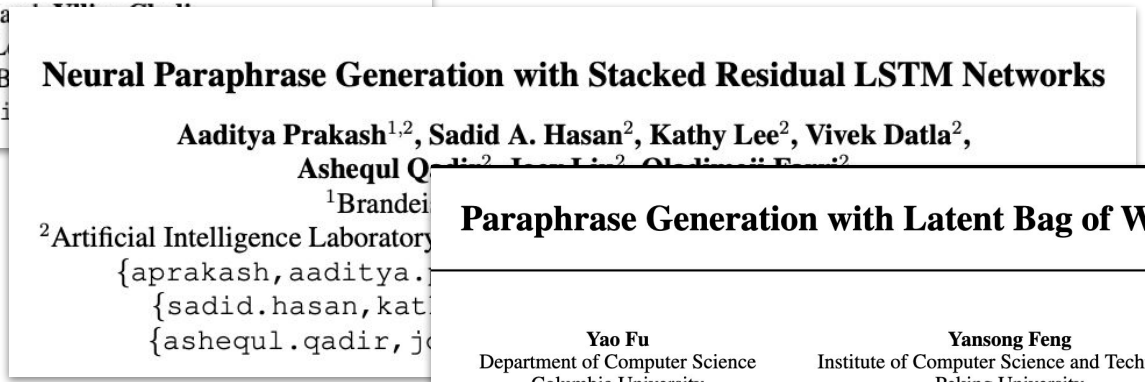
Who cares?

- Non-natives
  - Can use this as a lesson to improve their writing skills
  - Usually, a writing aims to be concise, however, when you want to explain something, you have to write more and more
- Students
  - Who want to get the good grade for their writing
- Researchers:
  - Who want to cite other research to their paper without plagiarism problem
- Journalist/writer
  - They usually need to write a lot to fill the empty spaces daily
  - So do people like to consume such content

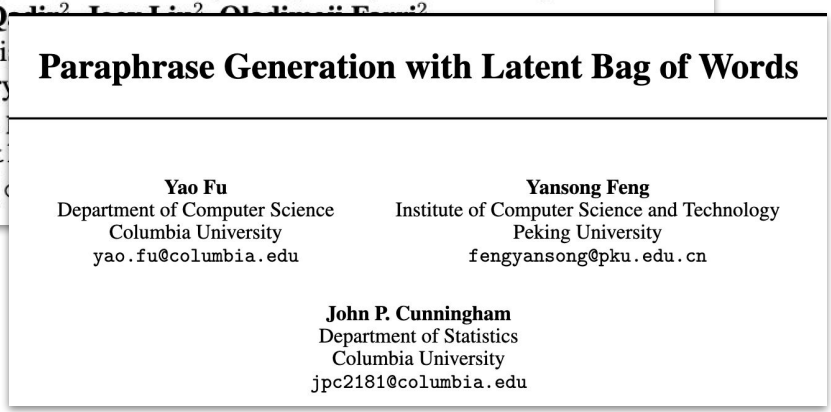
Current Work in the field



https://www.aclweb.org/anthology/D19-5627.pdf



https://arxiv.org/pdf/1610.03098.pdf



https://arxiv.org/pdf/2001.01941v1.pdf

What is new in our idea?

- Opposite to text summarization, we are expanding the text
- On top of paraphrasing which tries to rewrite sentences, we are focused on generating longer text
- Fairly different than current paraphrasing models
- Modifications in the beam search algorithm to be biased towards longer sentences

Dataset : Language-Net[1]

- A collection of sentence level paraphrases from Twitter
- The largest human-labeled paraphrase corpus to date of 51,524 sentence pairs

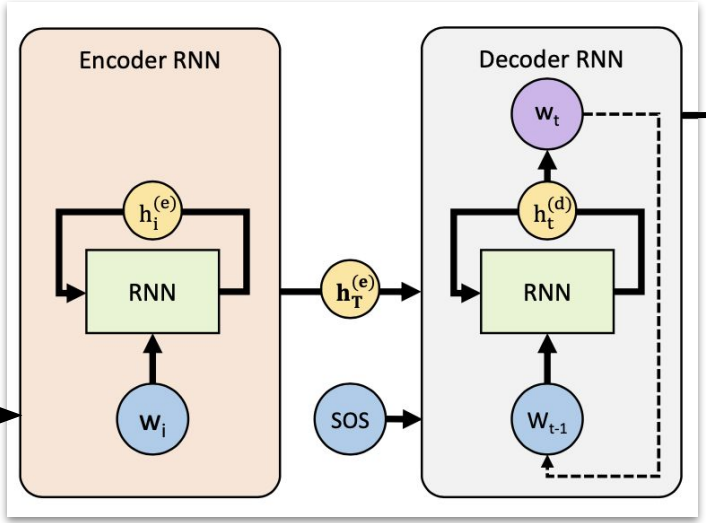
Sentence 1	Sentence 2	labeled
Samsung halts production of its Galaxy Note 7 as battery problems linger.	Samsung temporarily suspended production of its Galaxy Note7 devices following reports.	True
The 7 biggest changes Obamacare made, and those that may disappear.	What a repeal of Obamacare would look like , in plain English.	False

(1) https://lanwuwei.github.io/Twitter-URL-Corpus/

Encoder-Decoder LSTM

Sentence 1

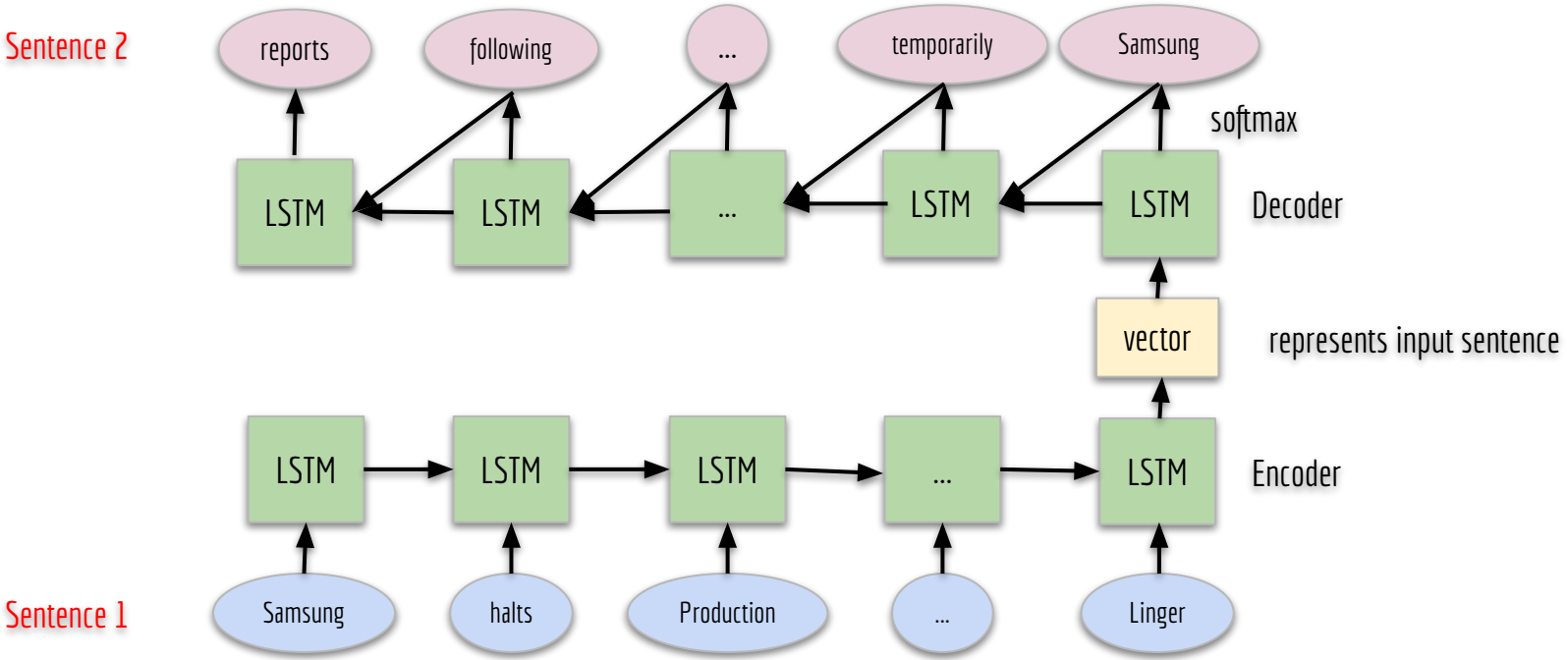
Samsung halts production of its Galaxy Note 7 as battery problems linger.



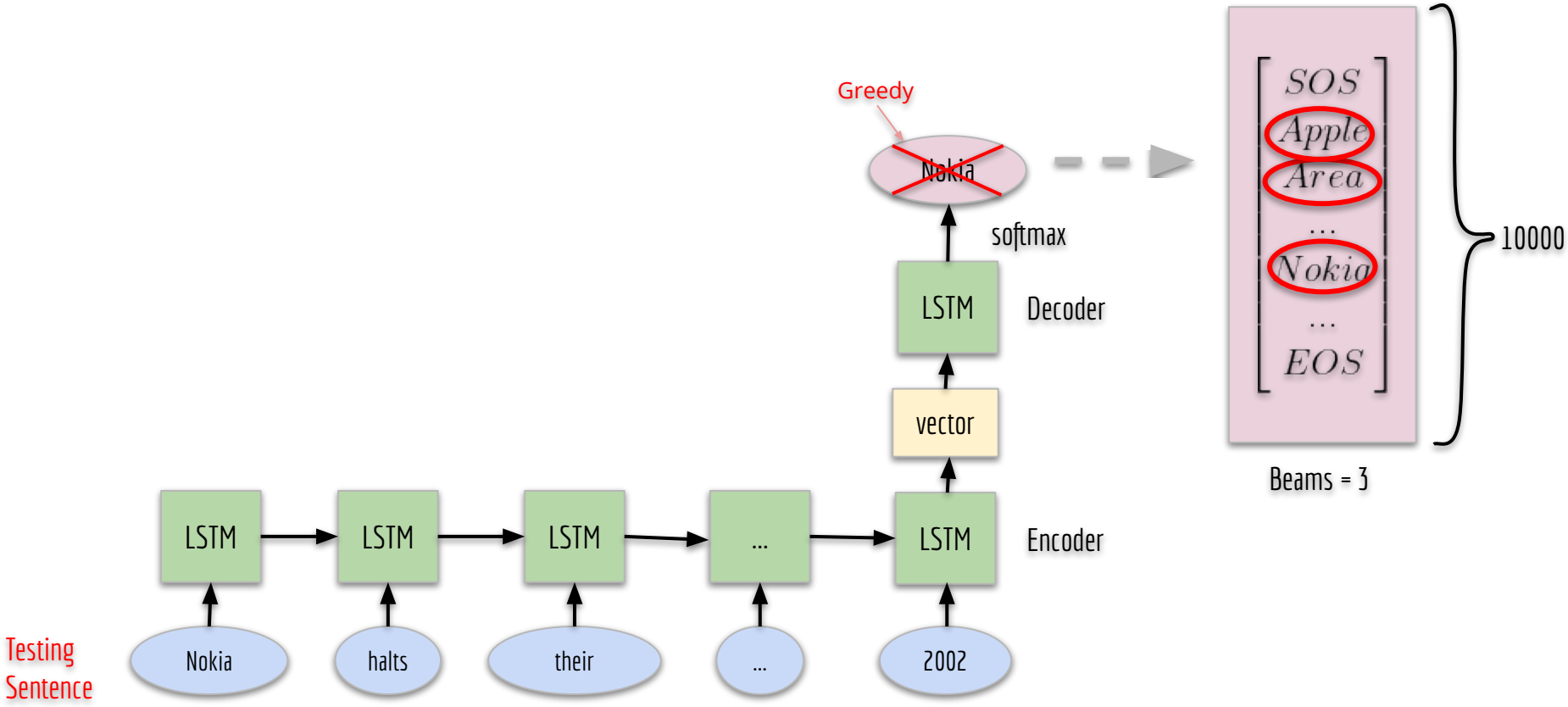
Sentence 2

Samsung temporarily suspended production of its Galaxy Note7 devices following reports.

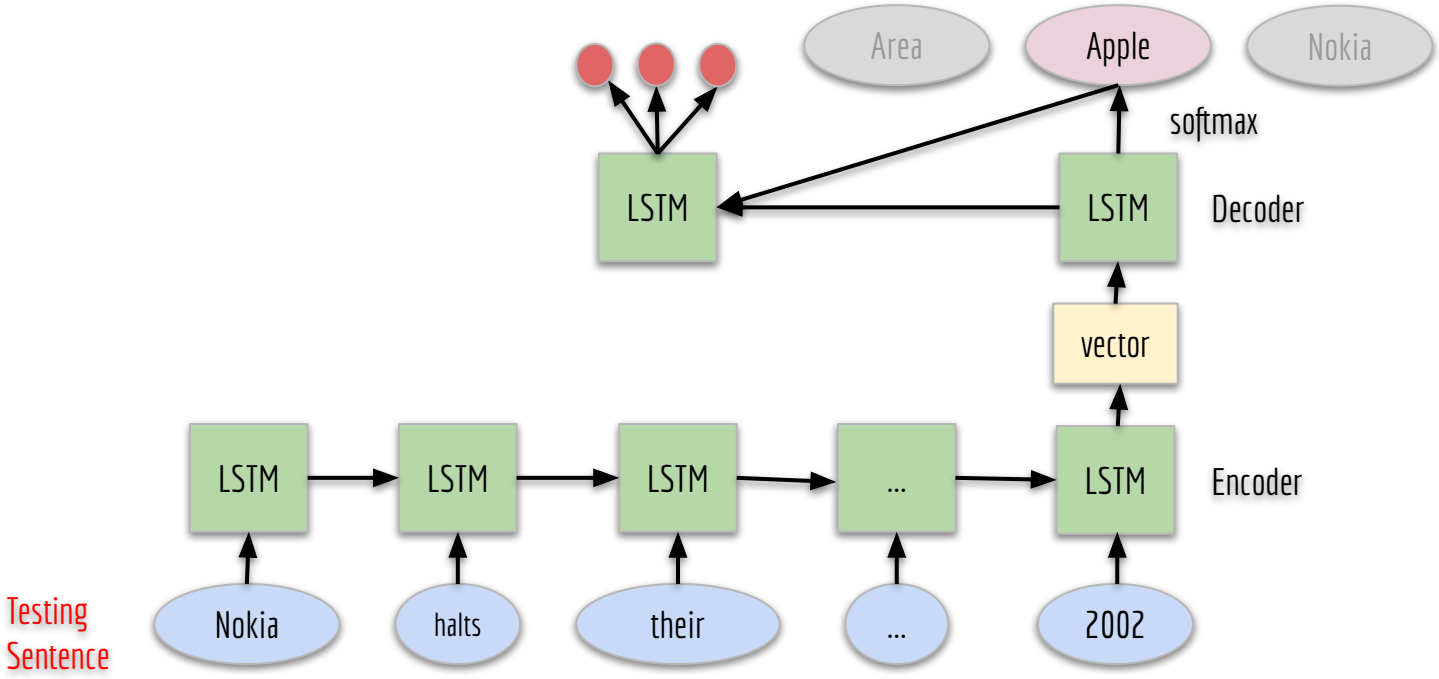
Encoder-Decoder LSTM



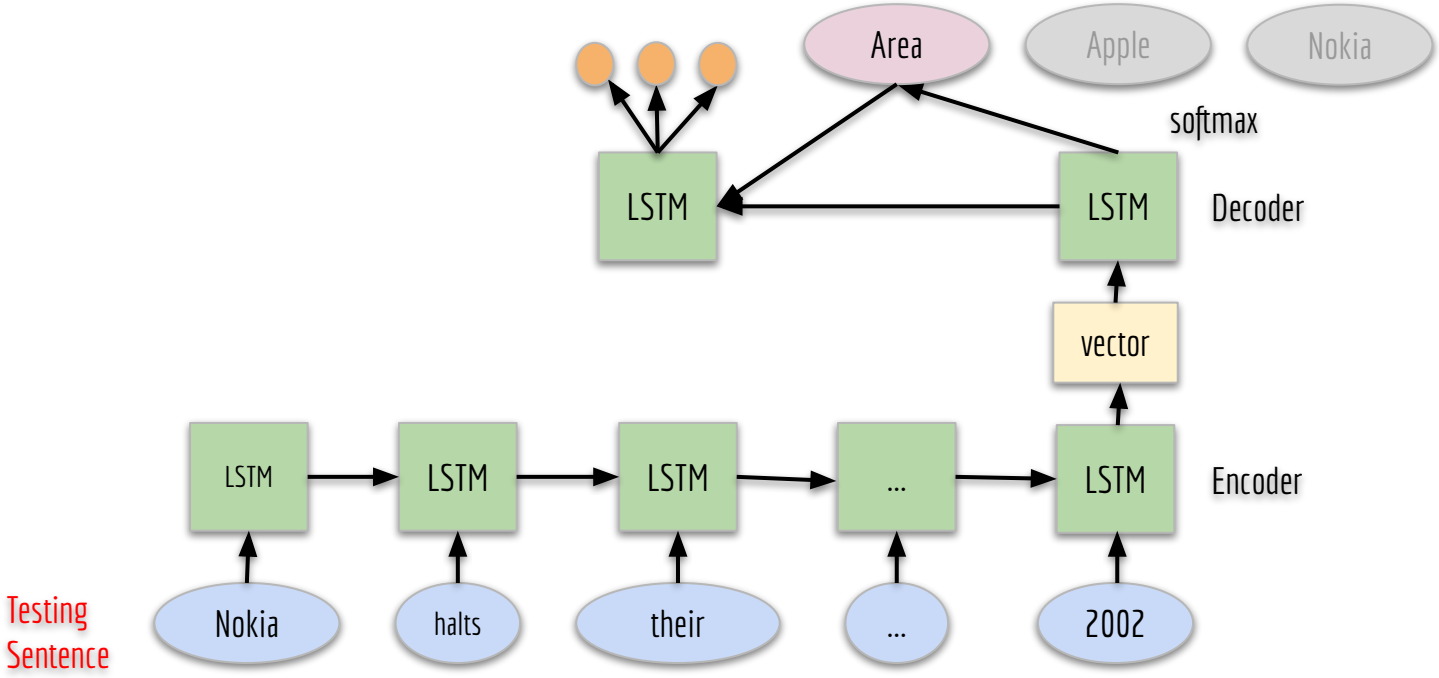
Encoder-Decoder LSTM + Beam Search



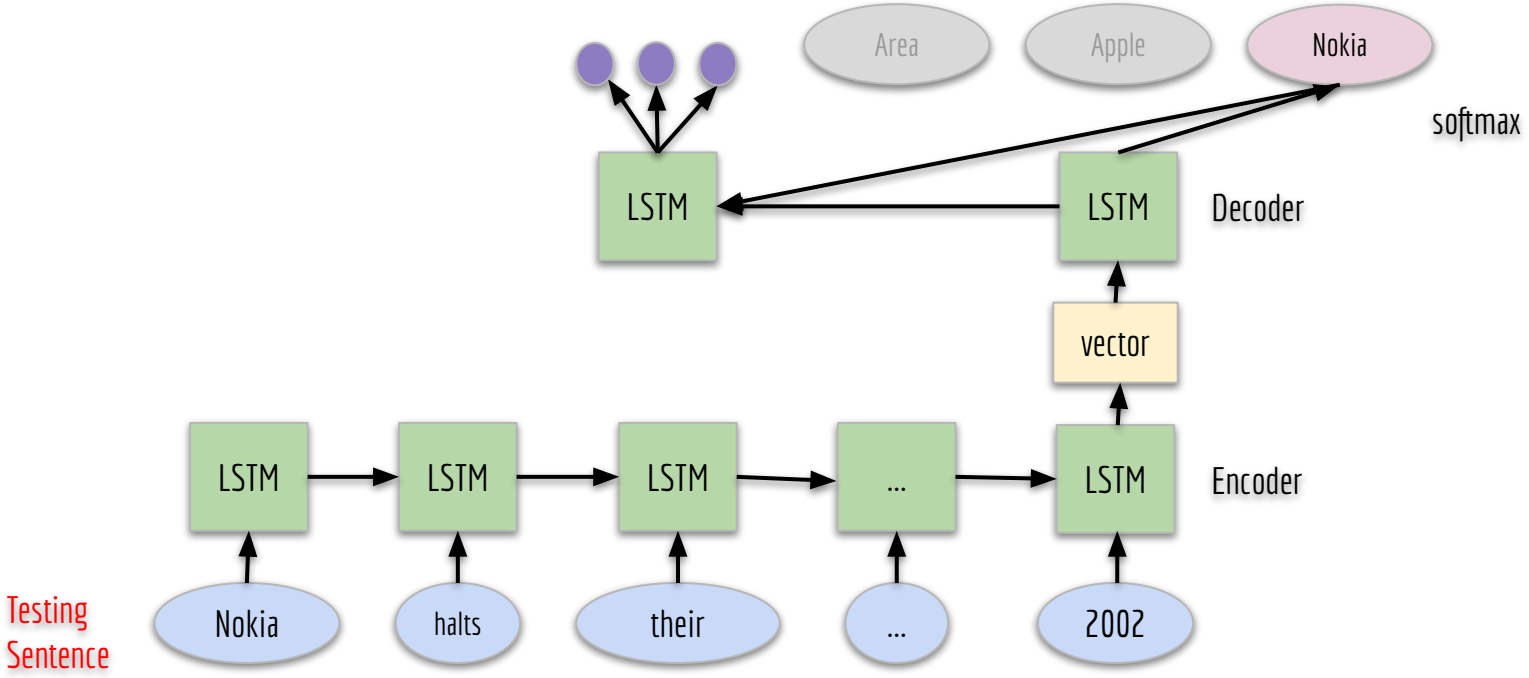
Encoder-Decoder LSTM + Beam Search



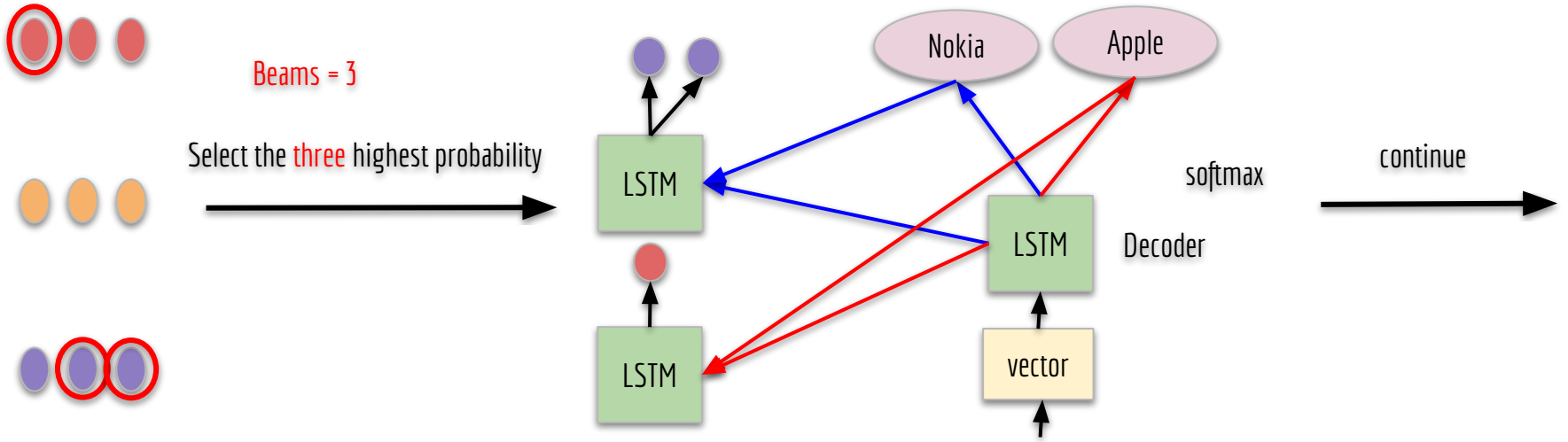
Encoder-Decoder LSTM + Beam Search



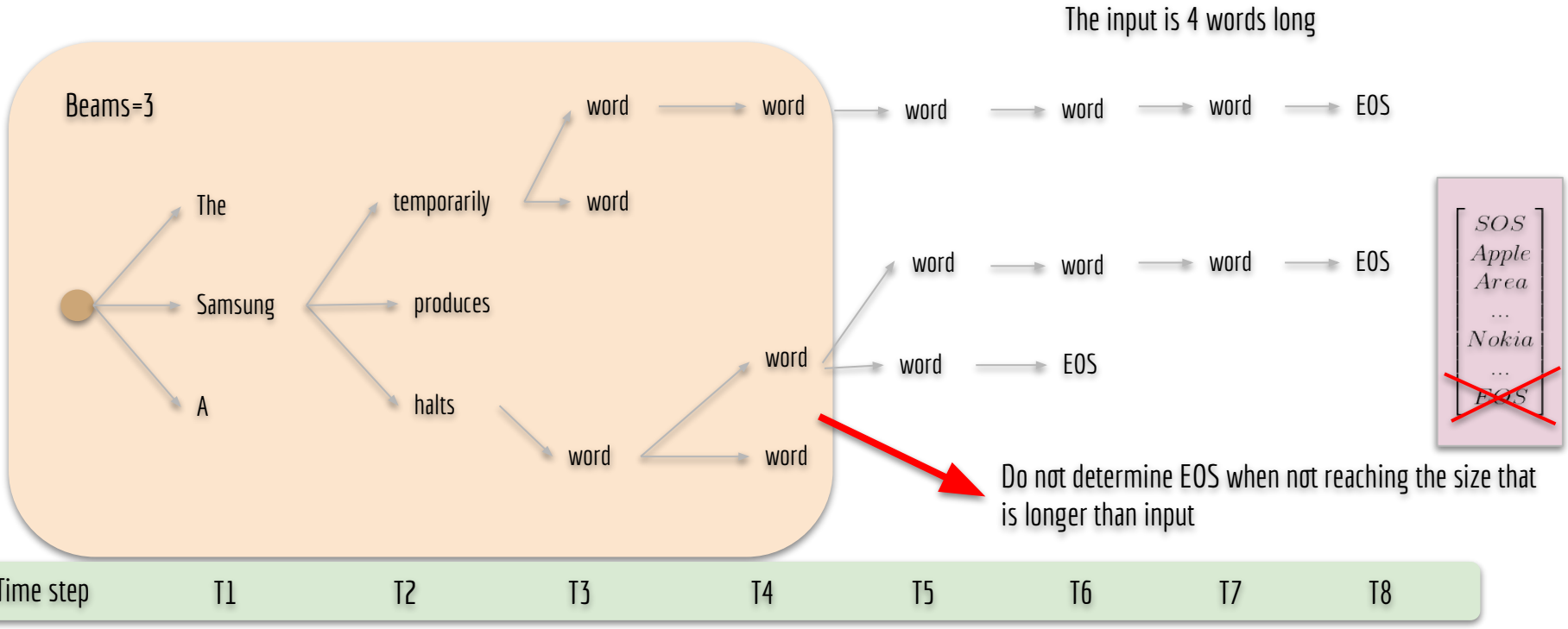
Encoder-Decoder LSTM + Beam Search



Encoder-Decoder LSTM + Beam Search



Little change for Beam Search



Project Risks

- Output sentence being:
  - Not paraphrased: adds excess details to the sentence or changes barely anything
  - Alters the sentence meaning even if it is longer than input
  - Not longer than the input: paraphrased but not longer
- **When we force the sentence to be longer (ignoring EOS in beam search if sentences don't reach the expected length), it might also break grammar/ sentence structure.**
- Too narrow/wide beam search can impact the quality of the sentences
- Ethical risk: model may be used for unethical purposes

Project Success

- Generate output sentences for test inputs
- Ask reviewers (this could be us, our friends, and family) to label the instances
  - Label 0: The generated sentence is NOT meaningful/paraphrased/longer
  - Label 1: The generated sentence conforms to project objective of generating a longer sentence whilst preserving its meaning
- Collect N reviewer's labels and pick the majority label

$$label_{out} = \begin{cases} 1, & \frac{\sum_{reviewer=1}^{N}(label_{reviewer} == 1)}{N} > 0.5 \\ 0, & Otherwise \end{cases}$$

- We'd like to achieve an accuracy (get more 1s than 0s) more than 70%

Project timeline

Week 7 - Week 8

- Build the model with dataset
- Evaluate the generated sentences

Week 9

- Experiment different models for encoder and decoder
- Modify the beam search further

Week 10

- Get reviewer labels on generated sentences
- Analyze model accuracies

Thanks!

=› We'd like to express our sincere gratitude to you for staying patient in listening to us today!