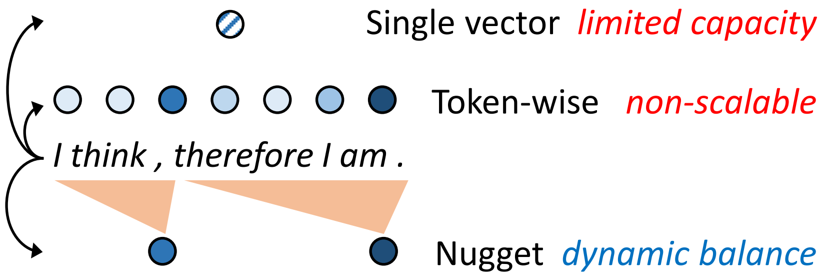
# **Nugget**: Neural Agglomerative Embeddings of Text

Guanghui Qin Benjamin Van Durme

JOHNS HOPKINS UNIVERSITY

## Cram sentence into a vector? No!



Single vector *limited capacity*

Token-wise *non-scalable*

*I think , therefore I am .*

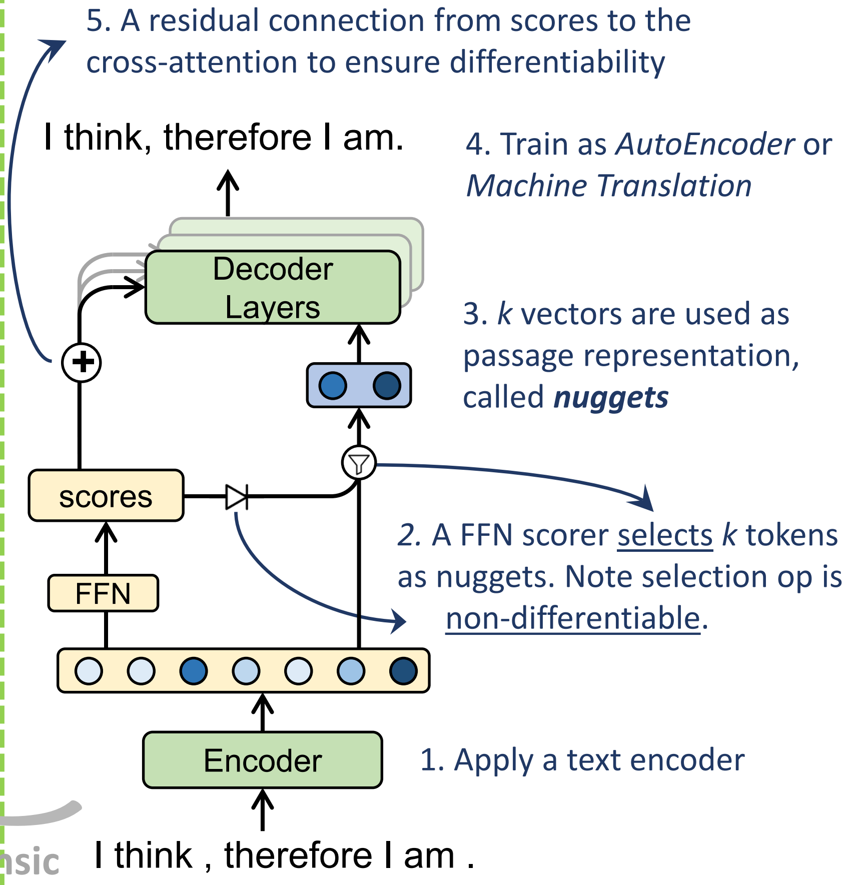
Nugget *dynamic balance*

Encode text with ***variable*** number of vectors!

Solution

## How?

*Key idea: Use the vectors of a **subset** of tokens to represent the whole passage*



5. A residual connection from scores to the cross-attention to ensure differentiability

4. Train as *AutoEncoder* or *Machine Translation*

3. *k* vectors are used as passage representation, called ***nuggets***

2. A FFN scorer selects *k* tokens as nuggets. Note selection op is non-differentiable.

1. Apply a text encoder

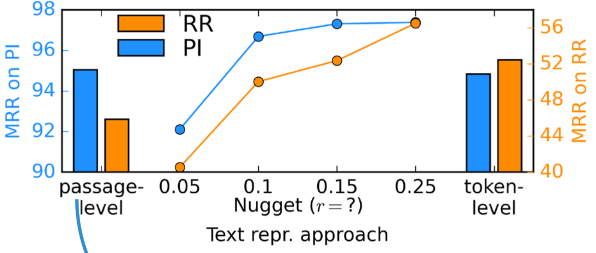
Extrinsic

## What are they good for?

### Document similarity test

Similarity between documents are measured by the many-to-many cosine similarity between their nugget vectors.

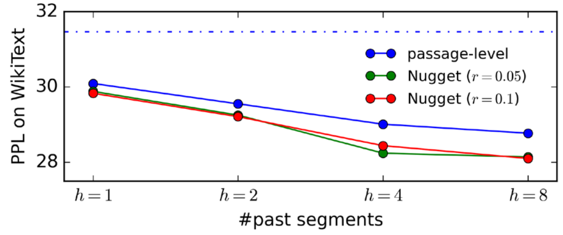
More nuggets -> better performance



Outperform the single-vector baseline

### LM with compressed context

Nugget is used to provide a compressed context to help a language model.



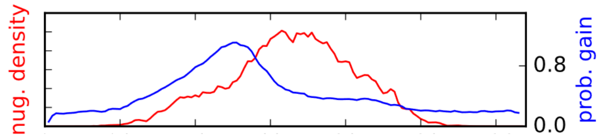
Nugget lowers the PPL. Longer compressed context consistently brings better performance.

## What are used as nuggets?

Run Nugget scorer and label nugget with boxes:

NLP is an interdisciplinary subfield of linguistics , computer science , and artificial intelligence concerned with the interactions between computers and human language , in particular how to program computers to process and …

Surprise? Nuggets are *text delimiters*!



And each nugget encodes *its preceding texts*!

Intrinsic