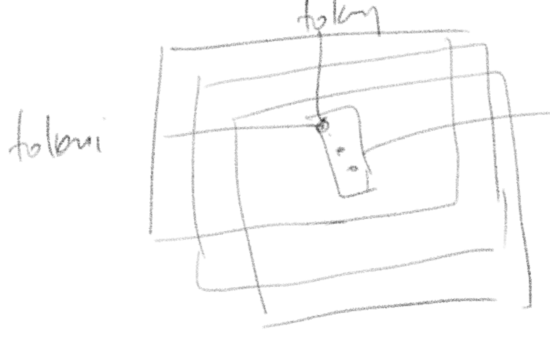
Exploring Geometry of BERT

Can understanding the geometry of internal representations tells us more about BERT, or help us improve BERT (optimize something that helps reduce the computation - mayb

- Does attention matrices encode syntactic features
- Dependency grammar relations
- Manning analyzed context encoding they are, using attention matrices.
- Attention probe -



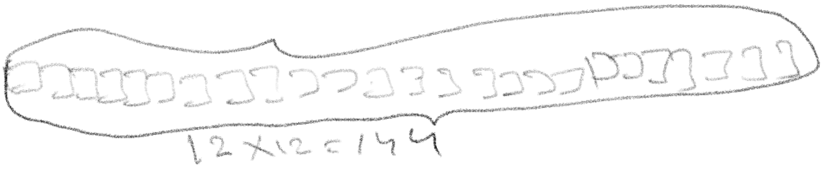
→ concat these to obtain attn vector ~~a_{ij}~~ a_{ij}

- every layer, every head

a_{ij} - ~~12 * 12~~

overlay

12 layer

single a_{ij} = 

Goal → classify a given relation b/w 2 tokens

↳ If good ⇒ modelwise attn vector ~~achieves~~ ~~be~~ encodes that relation.

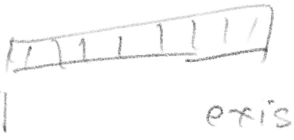
- 2 L2 linear models

① ↳ → dependency b/w i & j or not [binary classifier]

attn vector

↘ $\boxed{86.8\%}$

② multiclass classifier

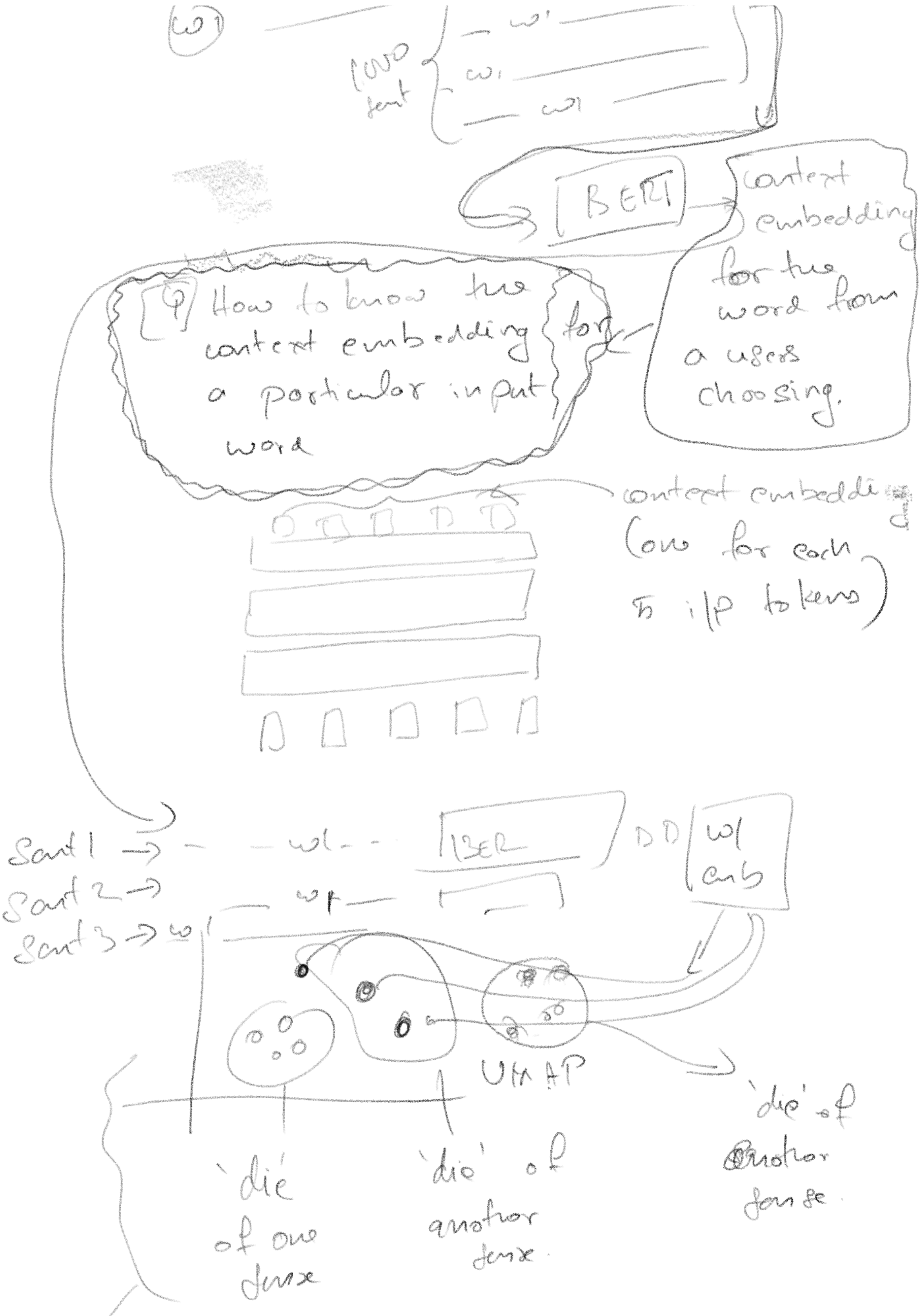
a_{ij}  → which ~~relat~~ dependancy exists b/w tokens i & j.

→ $\boxed{71.9\%}$

3.2 Geometry of parse tree embeddings

$$\boxed{\text{parse tree distance} = (\text{euclidean distance})^2}$$

4.1



w1
WSD
sent
w1
w1
w1
BERT
context embedding for the word from a users choosing.
Q How to know the context embedding for a particular input word
context embedding (one for each 5 i/p tokens)
Sent 1
w1
BER
w1 emb
Sent 2
w1
Sent 3
w1
UMAP
'die' of one sense
'die' of another sense.
'die' of another sense.

This technique even does WSD with

F1 = 71.1% (Nearest Neigh)

"[illegible] CAPTURING WORD SENSE" - mod

CONTEXT EMBEDDINGS ARE CAPTURING A WORD
sense due to "attention" - captures the context better.