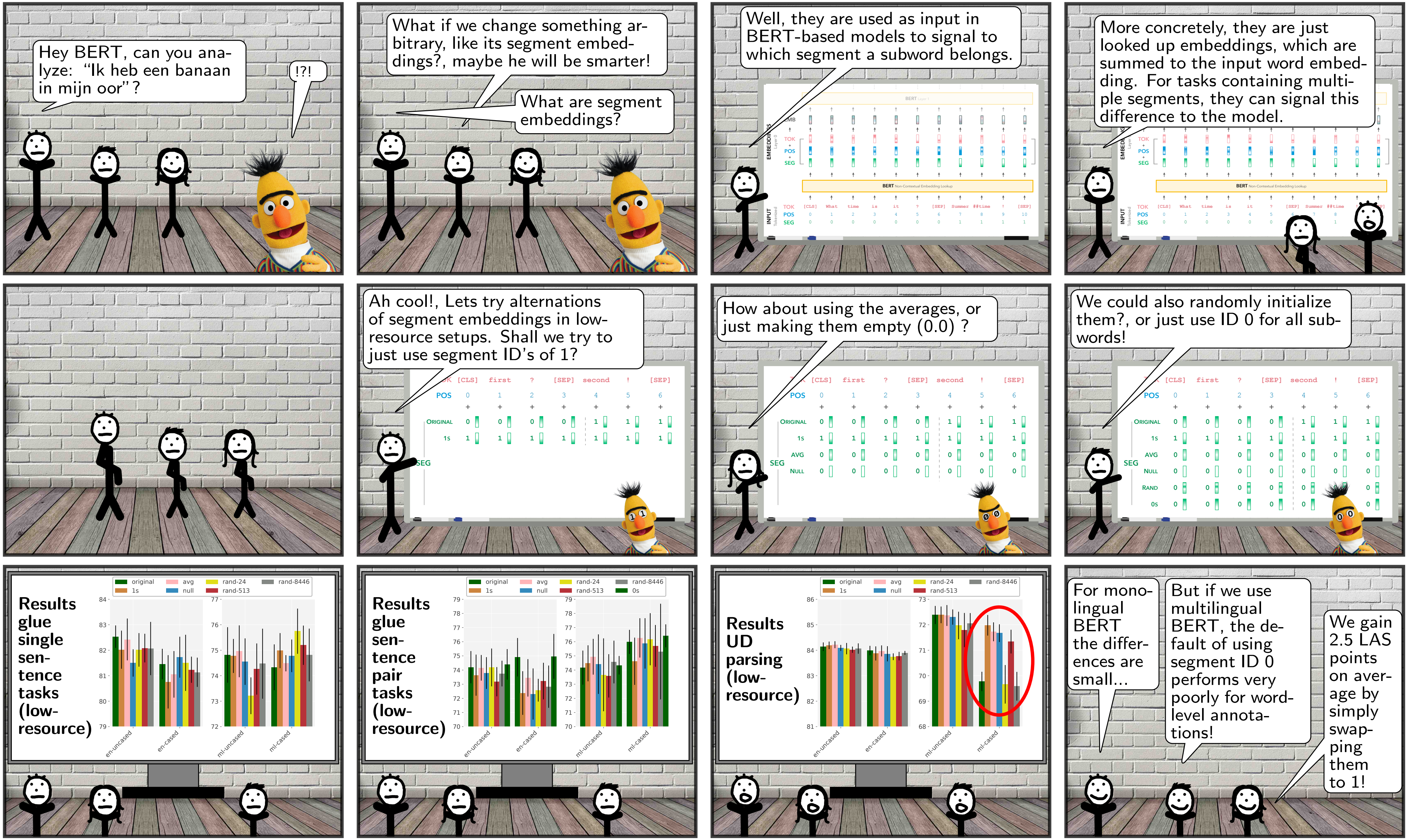
Frustratingly Easy Performance Improvements for Low-resource Setups: A Tale on BERT and Segment Embeddings

Rob van der Goot, Max Müller-Eberstein, Barbara Plank


Hey BERT, can you analyze: “Ik heb een banaan in mijn oor”?
!?!
What if we change something arbitrary, like its segment embeddings?, maybe he will be smarter!
What are segment embeddings?
Well, they are used as input in BERT-based models to signal to which segment a subword belongs.
EMBEDDINGS
INPUT
TOK
POS
SEG
[CLS] What time is it ? [SEP] Summer ##time [SEP]
More concretely, they are just looked up embeddings, which are summed to the input word embedding. For tasks containing multiple segments, they can signal this difference to the model.
Ah cool!, Lets try alternations of segment embeddings in low-resource setups. Shall we try to just use segment ID's of 1?
TOK [CLS] first ? [SEP] second ! [SEP]
POS 0 1 2 3 4 5 6
SEG
ORIGINAL 0 0 0 0 1 1 1
1s 1 1 1 1 1 1 1
How about using the averages, or just making them empty (0.0)?
AVG 0 0 0 0 0 0 0
NULL 0 0 0 0 0 0 0
We could also randomly initialize them?, or just use ID 0 for all subwords!
RAND 0 0 0 0 0 0 0
0s 0 0 0 0 0 0 0
Results glue single sentence tasks (low-resource)
original
1s
avg
null
rand-24
rand-513
rand-8446
84
83
82
81
80
79
77
76
75
74
73
72
en-uncased
en-cased
ml-uncased
ml-cased
Results glue sentence pair tasks (low-resource)
original
1s
avg
null
rand-24
rand-513
rand-8446
0s
79
78
77
76
75
74
73
72
71
70
en-uncased
en-cased
ml-uncased
ml-cased
Results UD parsing (low-resource)
original
1s
avg
null
rand-24
rand-513
rand-8446
86
85
84
83
82
81
73
72
71
70
69
68
en-uncased
en-cased
ml-uncased
ml-cased
For monolingual BERT the differences are small...
But if we use multilingual BERT, the default of using segment ID 0 performs very poorly for word-level annotations!
We gain 2.5 LAS points on average by simply swapping them to 1!

*This is a work of fiction. Any similarities to persons or actual events is purely coincidental.