Dialectal Extractive Question Answer

Adel Alhamisy, Aniket Pandey, Pankaj Jatav, Yash Bagal

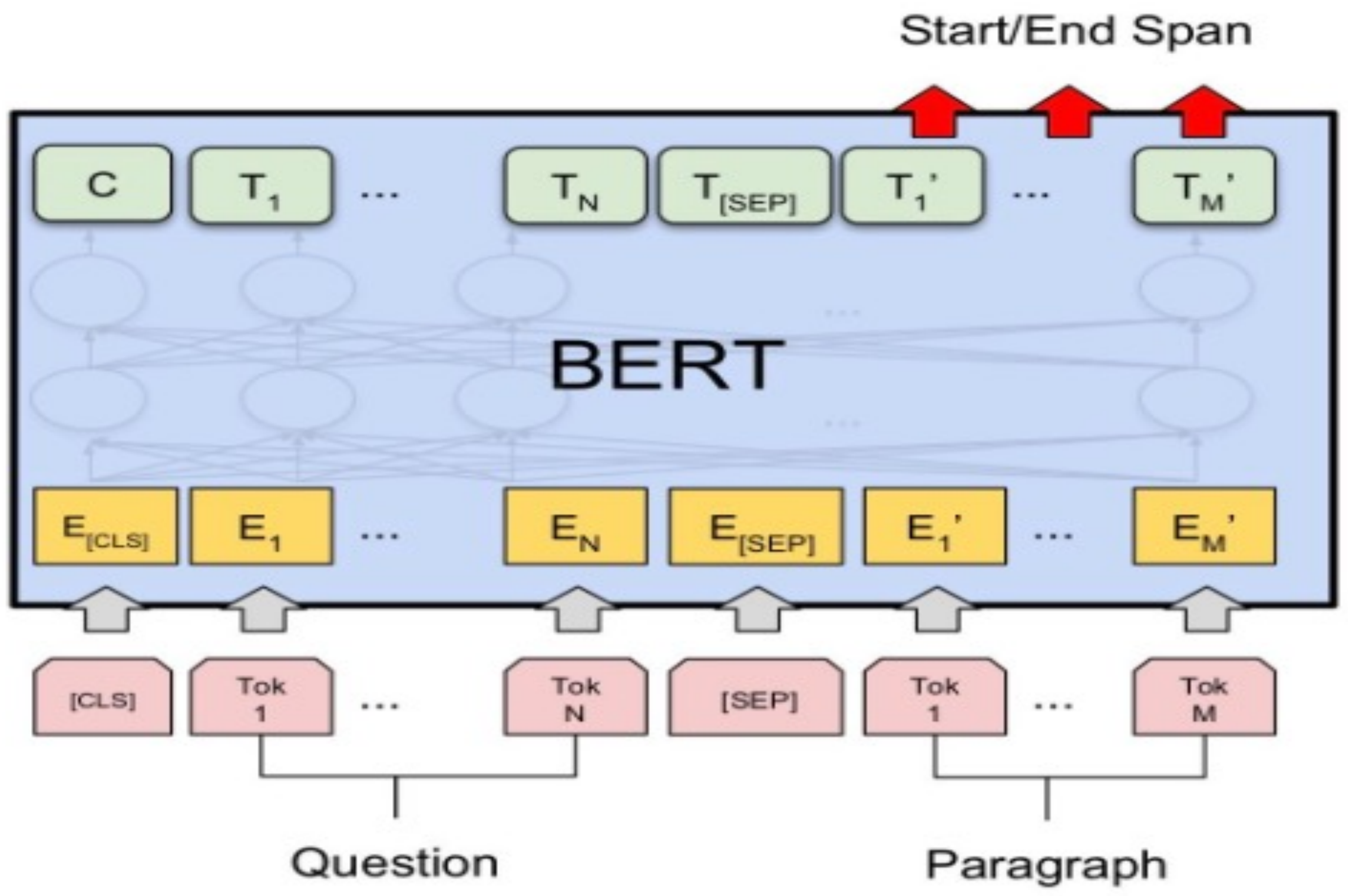
{aalkhami, apandey7, pjatav, ybagal}@gmu.edu

GEORGE MASON UNIVERSITY

Abstract

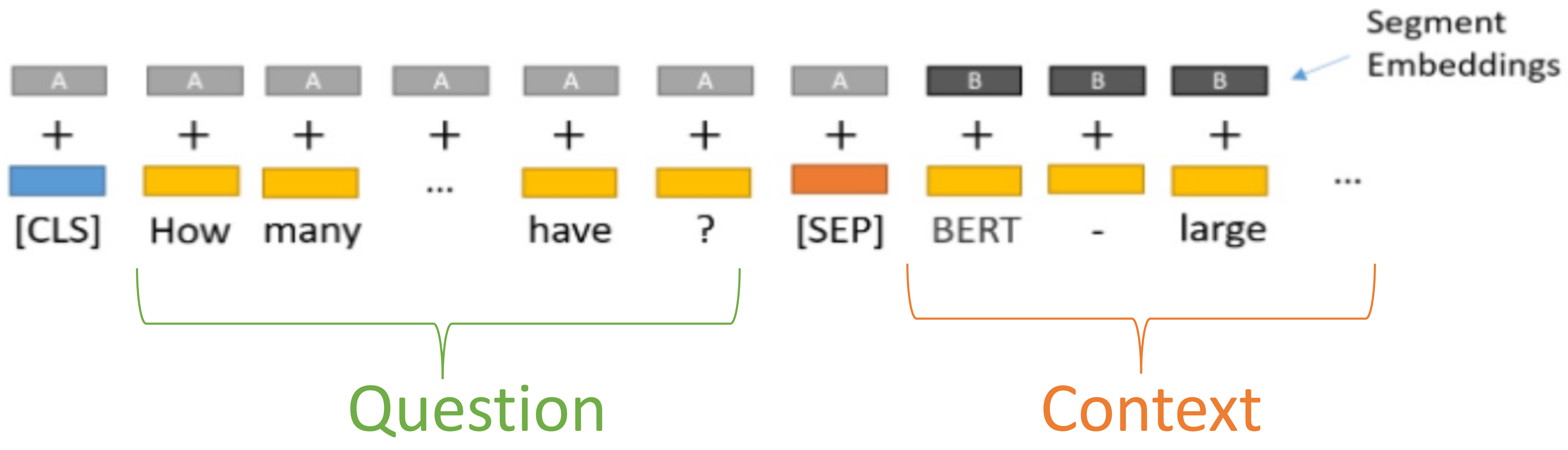
We build a QA system that can recognize different languages dialects and output an answer in response to a question. The current QA systems do not consider the faults that might be introduced by different dialects. We propose a BERT with Language-Clustered vocabulary (Chung et al., 2020) or Poolingformer (Zhang et al., 2021) in an effort to improve the performance.

Architecture



Dataset – 5 Languages, 24 Dialects

Language	Dialect (Variety Code)
Arabic	Algeria (DZA), Bahrain (BHR), Egypt (EGY), Jordan (JOR), Morocco (MAR), Saudi Arabia (SAU), Tunisia (TUN)
Bengali	Bangladesh-Dhaka (BGD), India- Kolkata (IND)
English	Australia (AUS), India-South (IND-S), India-North (IND-N), Ireland (IRL), Kenya (KEN), New Zealand (NZL), Nigeria (NGA), Philippines (PHI), Scotland (SCO), South Africa (ZAF), US-Southeast (USA-SE)
Korean	South Korea-Seoul (KOR-C), South Korea-south (KOR-SE)
Kiswahili	Kenya (KEN), Tanzania (TZA)



Question: How many parameters does BERT-large have?

Context: BERT large is really big… it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.24GB, so expect it to take couple of minutes to load on your instance.

Answer: 340M

Question: من قام بلعب دور هاري بوتر في الأفلام المصورة؟

Context: دانيال رادكليف في دور هاري بوتر ، وهو يتيم يبلغ من العمر 12 عامًا ، يقيم عند عمته وعمه الذي لا يرحم ، مشهور بسبب أنه نجا من محاولة قتل على يد الساحر المظلم لورد فولدمورت عندما كان رضيعا، لكن نجح فولدمورت في قتل أبويه ،و هو طالب في مدرسة هوجورتس للسحر[1][1]

Answer: دانيال رادكليف

Analysis / Experiments

- The ffaisal93/DialQA uses bert-base-multilingual-uncased which has normalization issue in many languages and therefore according to the official BERT documentation it is not recommended so, we switched to bert-base-multilingual-cased model which comprises of 104 languages.
- Using the bert-base-multilingual-uncased we got an average accuracy of 70% on all languages.
- We are working on using Poolingformer which beats previous state of the art model TyDi QA by 1.9 points.

CS-678: Advanced NLP
Department of Computer Science
School of Computing