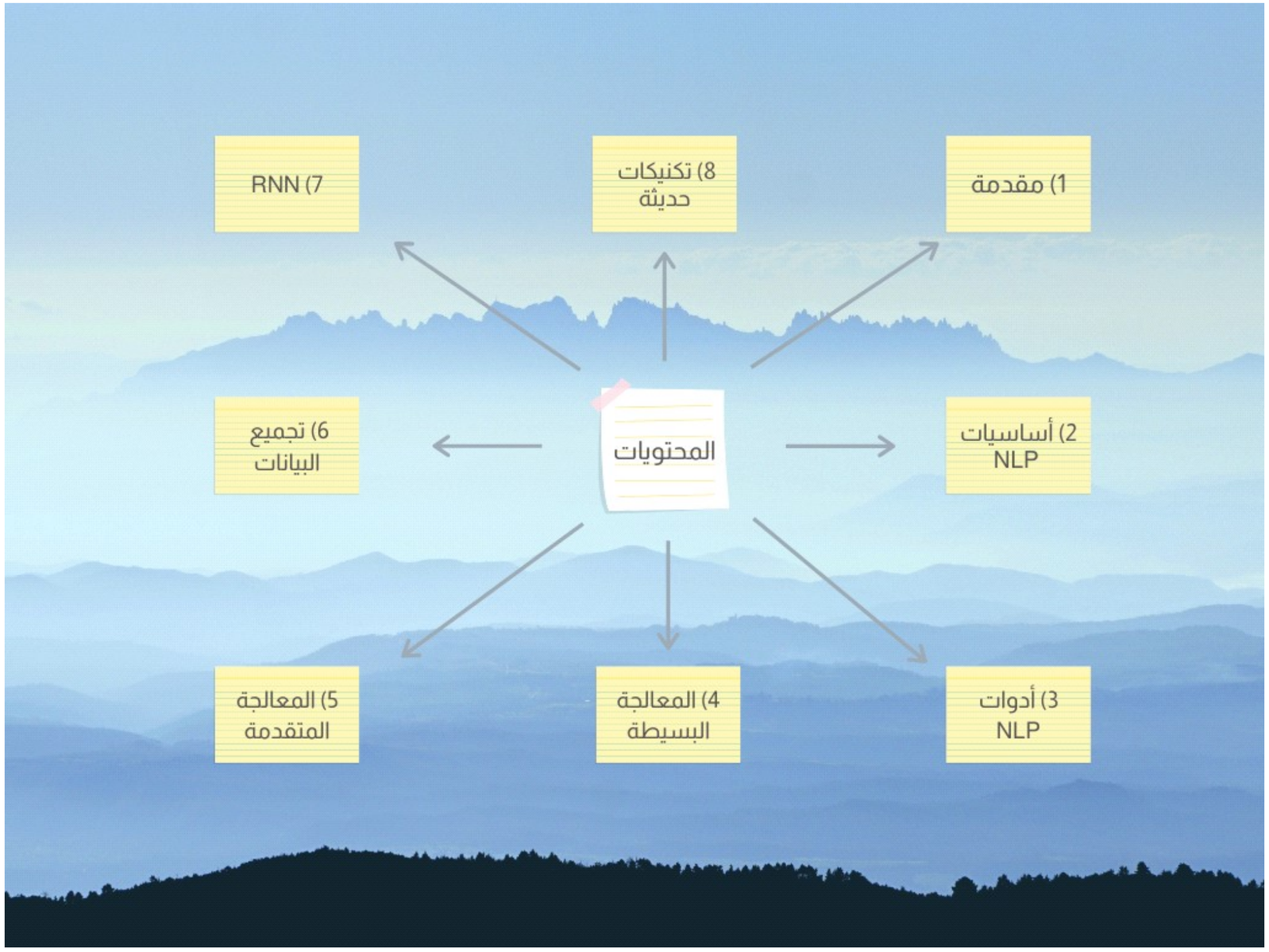
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

				التطبيقات	العقبات و التحديات	تاريخ NLP	ما هو NLP	المحتويات	1) مقدمة
					البحث في النصوص	ملفات pdf	الملفات النصية	المكتبات	2) أساسيات NLP
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	3) أدوات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning	4)المعالجة البسيطة
T. Generation	L. Modeling	NGrams	Lexicons	GloVe	NMF	LDA	T. Clustering	T. Classification	5)المعاجلة المتقدمة
	Summarization & Snippets		Ans. Questions		Auto Correct	Vader	Naïve Bayes	Sent. Analysis	
Search Engine	Relative Extraction		Information Retrieval		Information Extraction		Data Scraping	Tweet Collecting	6) تجميع البيانات
					Rec NN\TNN	GRU	LSTM	RNN	RNN (7
Chat Bot	Gensim	FastText	Bert	Hug. Face	Attention Model	T. Forcing	CNN	Word Cloud	8) تكنيكات حديثة

القسم الثامن : تكنيكات حديثة

الجزء السادس : Bert

==========

نتناول الآن أحد أحدث إصدارات جوجل , وهو الالجوريثم الذي تم تدريبه بالفعل BERT و هو اختصار (Bidirectional Encoder Representations from Transformers) و الذي تم ابتكاره في 2018 , واستخدامه في 2019 , و هو من تصميم كلا من (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova)

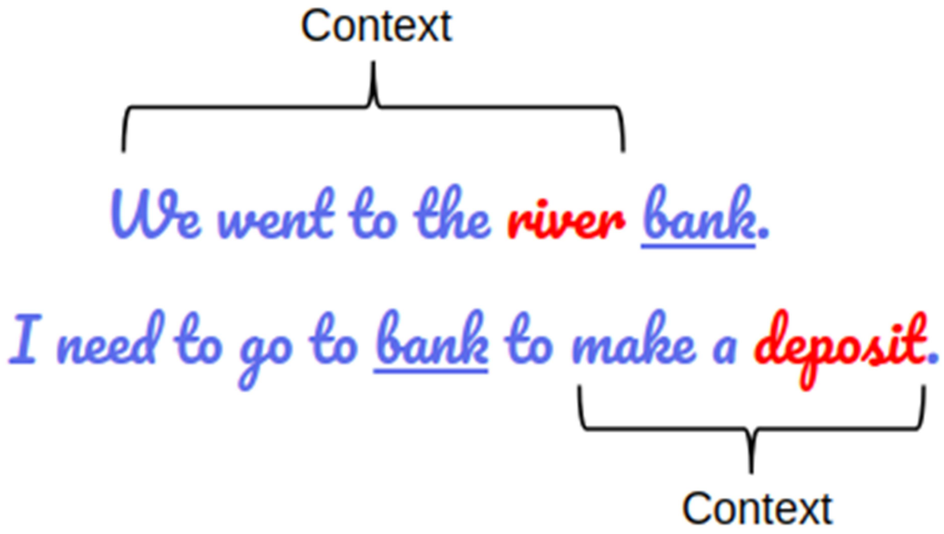
و فكرة الخوارزم ,معتمدة علي نجاح جوجل في تدريبها الموديل علي عشرات المليارات من الجمل , لتمييز الجمل الصحيحة عن الخطأ , وذلك للتركيز علي الاسلوب الصحيح لصياغة الجمل , حتي ان تم تدريبها علي ويكيبيديا بالكامل , بالإضافة الي عشرات الالف من الكتب, بمجموع قريبا 3.3 مليار كلمة .

و يستخدم هذا الخوارزم بشكل اساسي للتعامل مع محرك بحث جوجل , وهو ما صنع طفرة كبيرة في امكانية البحث , و جعل جوجل قادرة علي فهم ما الذي تريده من خلال كلمات قليلة , واختيار النتائج المطلوبة بالتحديد بشكل اكثر دقة

فنموذج BERT يتناول كل كلمة علي حدة , ويفهم منها المعني الدقيق , دون ان يعرض لك نتائج اخري بعيدة عن الشئ المطلوب , و اي مستخدم لجوجل قوي الملاحظة , سيجد ان نتائج البحث في الفترة الاخيرة صارت اكثر دقة ,في اختيار النتائج

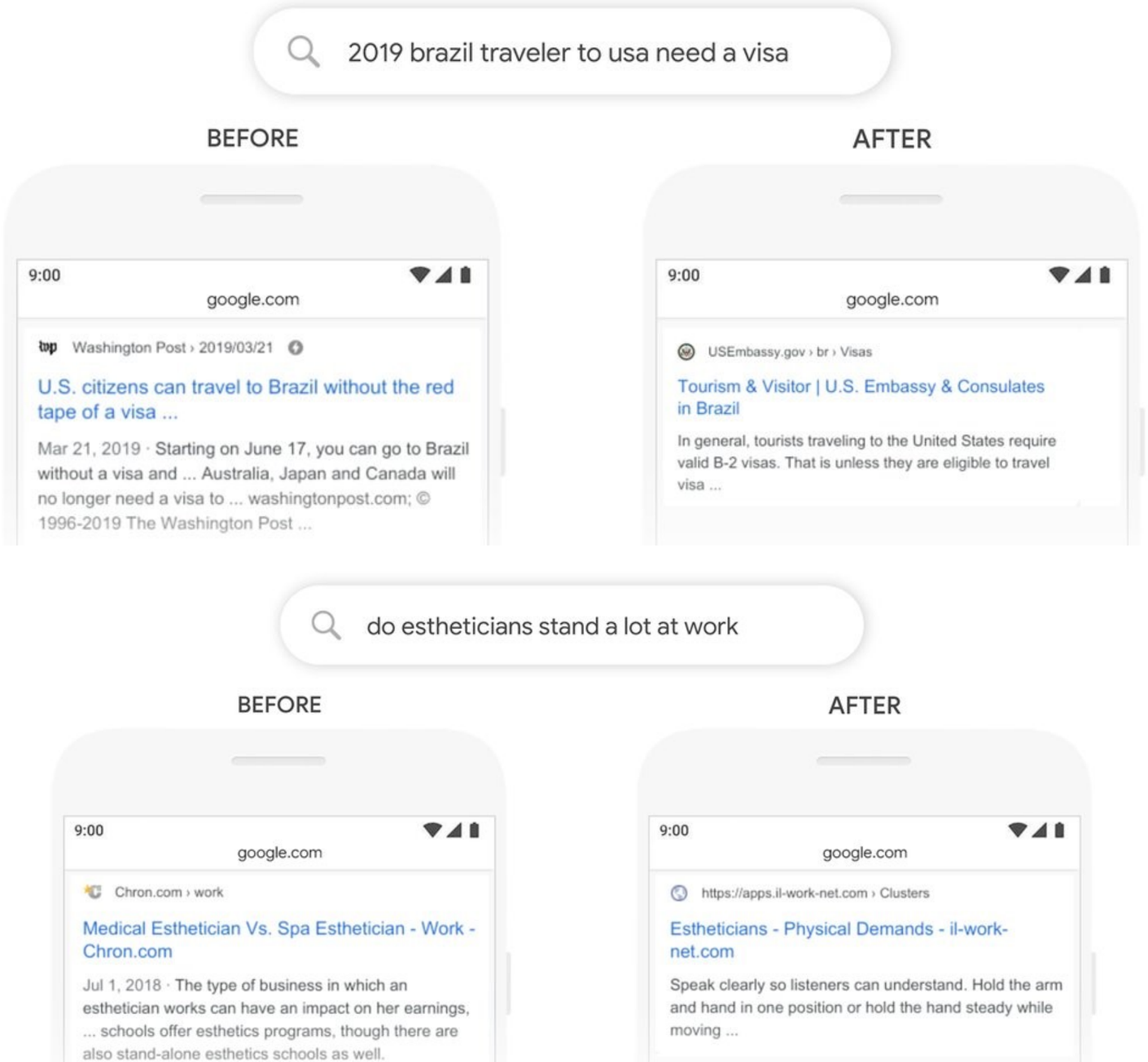
و لأن BERT هو يعتمد علي bidirectional encoder فهو يتناول النصوص من الجانبين , كما رأينا في BRNN

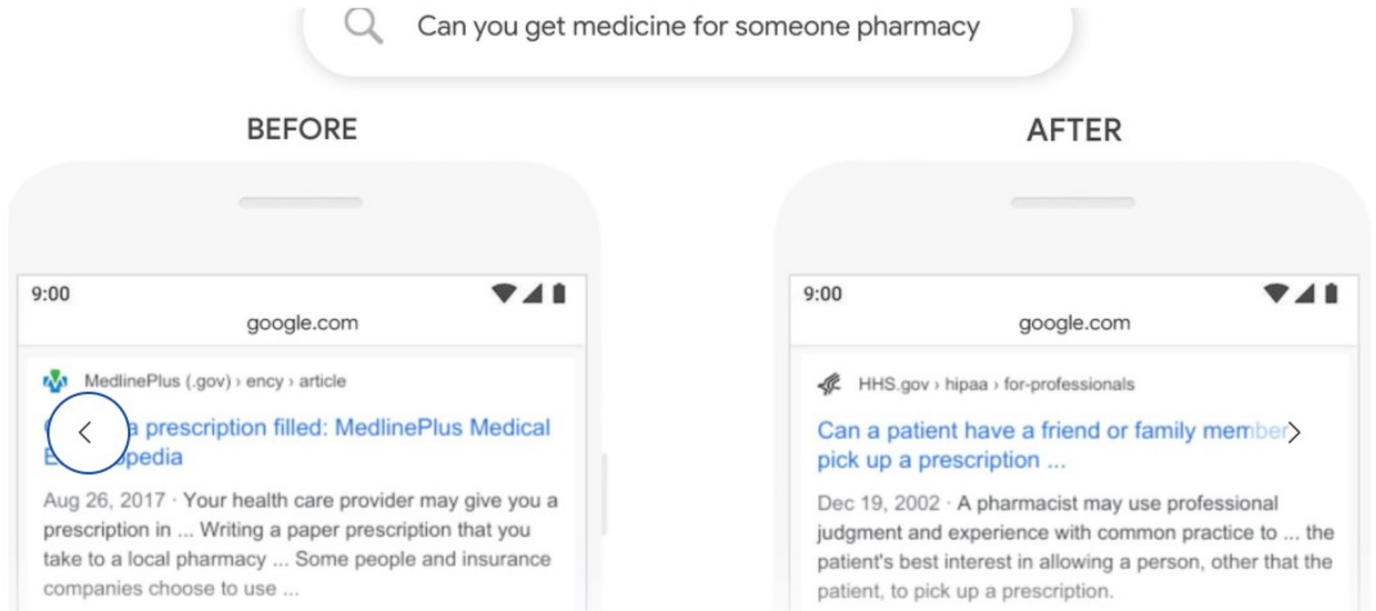
و أهمية التدريب من اتجاهين تظهر في المثال التالي :



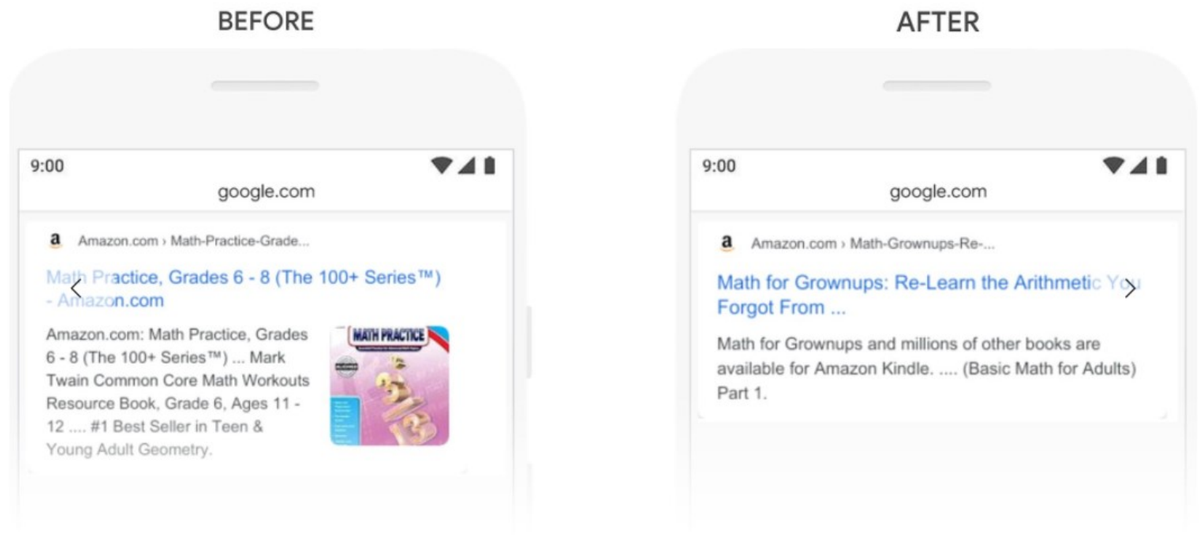
فمعني كلمة bank يختلف في الجملتين , بناء علي الكلمة السابقة لها في الجملة الاولي , و التالية لها في الجملة الثانية

و هنا عدد من الامثلة التي اوردتها جوجل لاستخدامات بيرت

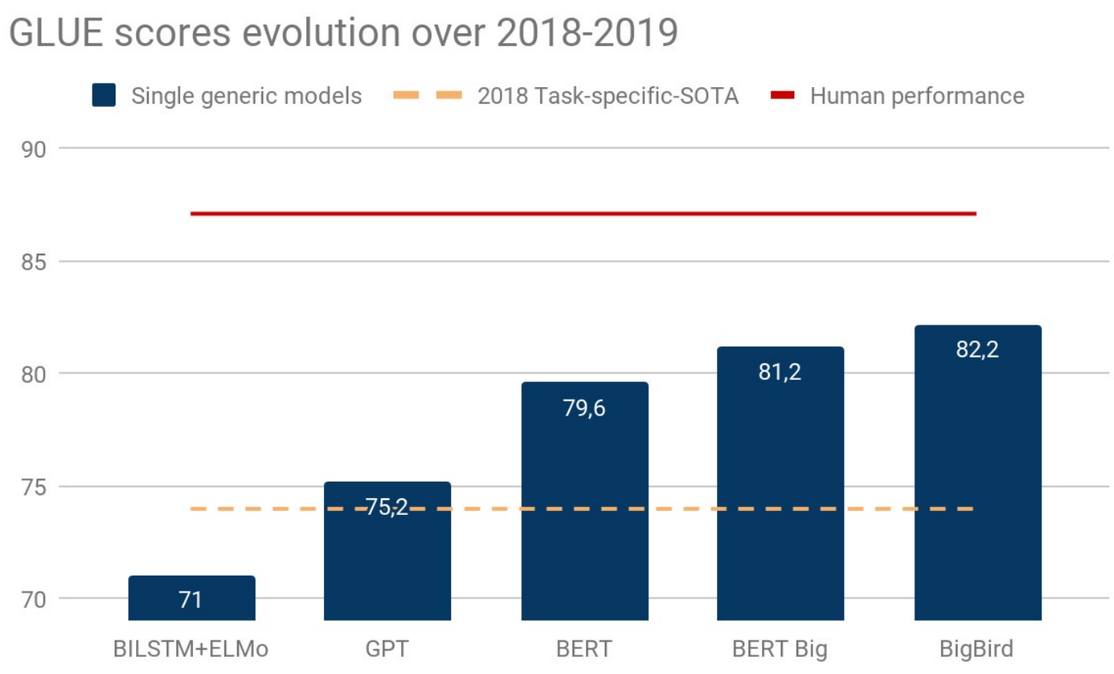




In the past, a query like this would confuse our systems--we placed too much importance on the word "curb" and ignored the word "no", not



و هذا الجراف, يظهر كيف ان bert & bigbird اقتربوا كثيرا من سقف الدقة البشرية خلال عام واحد



و هذا ليس الإصدار الأول لجوجل في مجال استخدام خوارزم لمعالجة البحث , فقبل ذلك كام يستخدم عدد من الخوارزميات مثل : RankBrain , Panda , Penguin و هكذا , ويتوقع ان يقوم BERT بالتعاون مع RankBrain و ليس لاحلال مكانه

و قد ذكر جاكوب ديفلين , ان بيرت قادر علي توقع الكلمات الناقصة بنسبة كبيرة من النجاح , فعلي موقع googleblog ذكر المثال التالي :

Input: The man went to the $[MASK]_1$. He bought a $[MASK]_2$ of milk .
Labels: $[MASK]_1$ = store; $[MASK]_2$ = gallon

ايضا بيرت قادر علي تمييز هل الجمل متعاقبة في المعني ام ليس لها علاقة ببعضها البعض :

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

كما استعرض عددا من الإحصائيات و الدقة لمقارنة بيرت بخوارزيمات اخري بل و بدقة الإنسان

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

و قد قامو بجعل الكود الكامل له متاح مجانا علي جيتهاب من هنا

https://github.com/google-research/bert

و هذا هو الوصف الذي قام مصممو BERT بكتابته علي موقع جامعة كورنيل :

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers.

Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

و هنا وصف مناسب عنه

BERT stands for Bidirectional Encoder Representations from Transformers. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks

و بالتالي تكون الخطوتين الاساسيتين للتعامل مع بيرت هو :

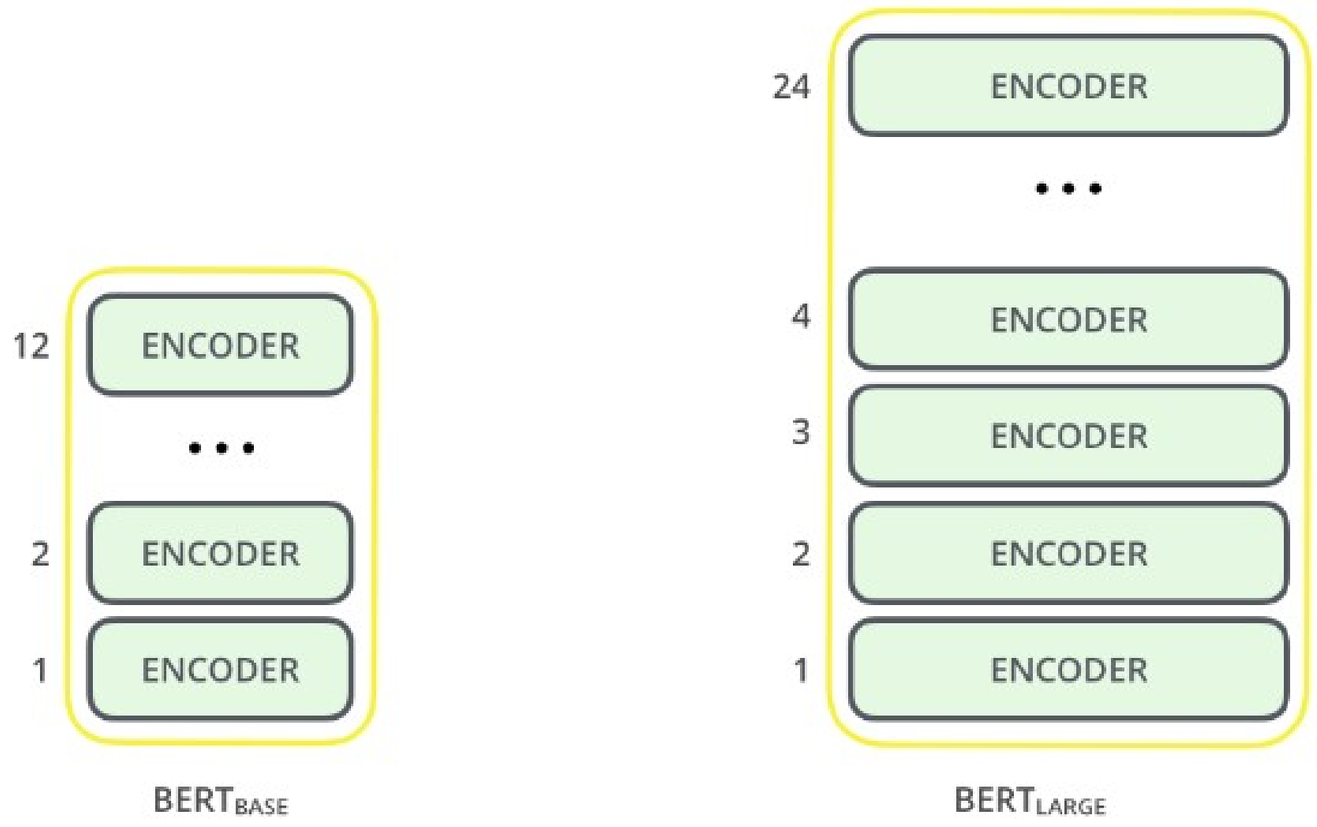
1. التدريب الضخم الذي يتم علي كمية هائلة من البيانات الغير معنونة (التدريب دون إشراف) وهو ما تم بالفعل و يتم باستمرار في معامل جوجل
2. التدريب الأخير علي الداتا الخاصة بك (التدريب بإشراف) , لتكون جاهزة لتطبيقها

و هذا الأمر يعني اننا نقوم بنوع من Transfer learning , حينما نأتي بخوارزم تم تدريبه , ونقوم بتدريبه بشكل نهائي علي البيانات الخاصة بنا

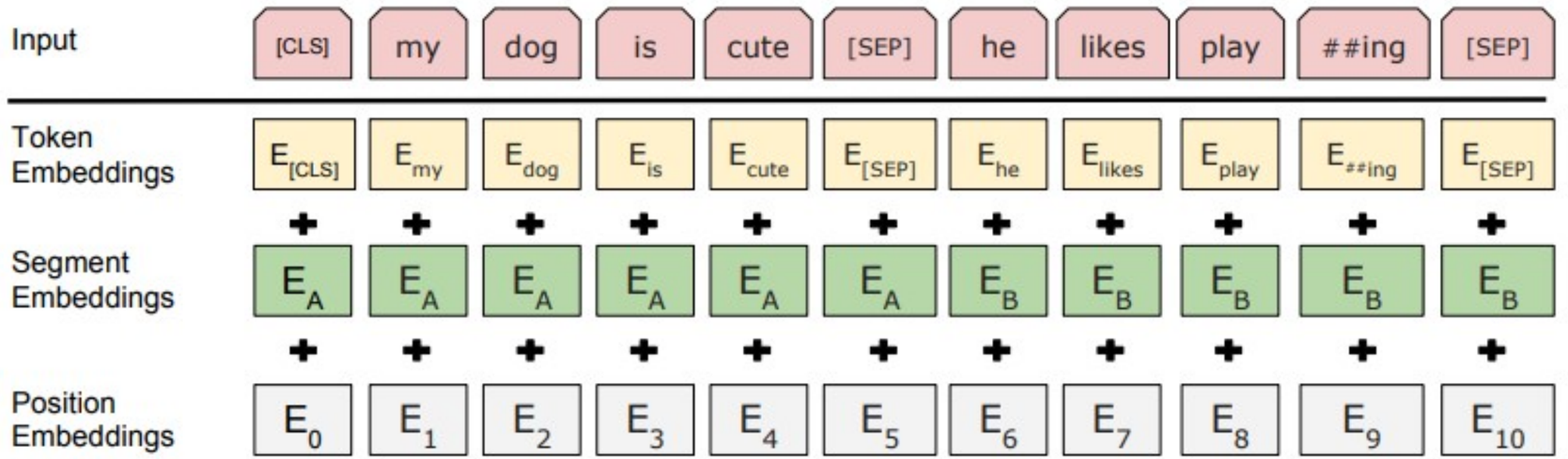
> *Transfer Learning in NLP = Pre-Training and Fine-Tuning*

و هنا تصميمات من الهيكل الداخلي له :

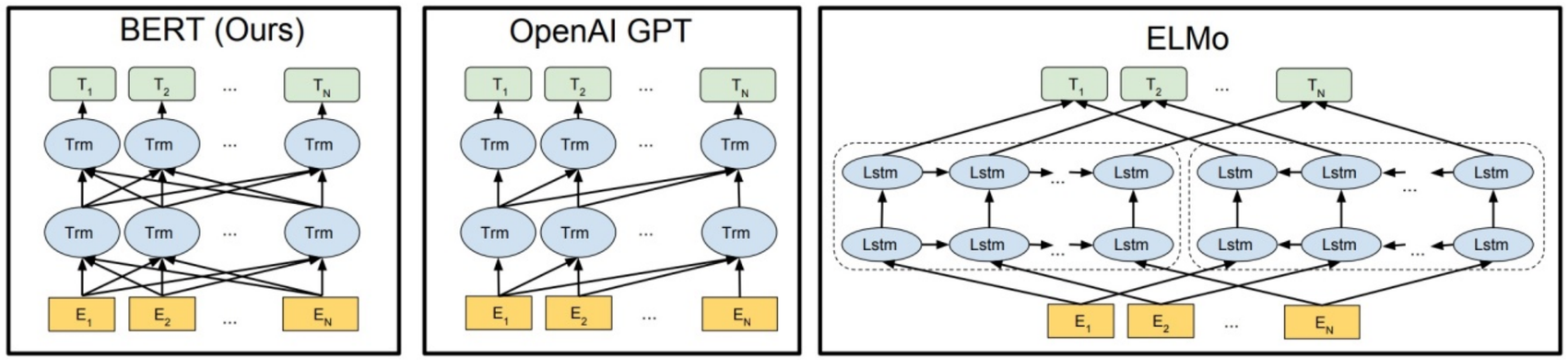
فالنموذج العادي منه bert base يتكون من 12 طبقة بعدد 110 مليون باراميتر
اما النموذج الأكبر , فهو يتكون من 24 طبقة , بعدد 340 مليون باراميتر



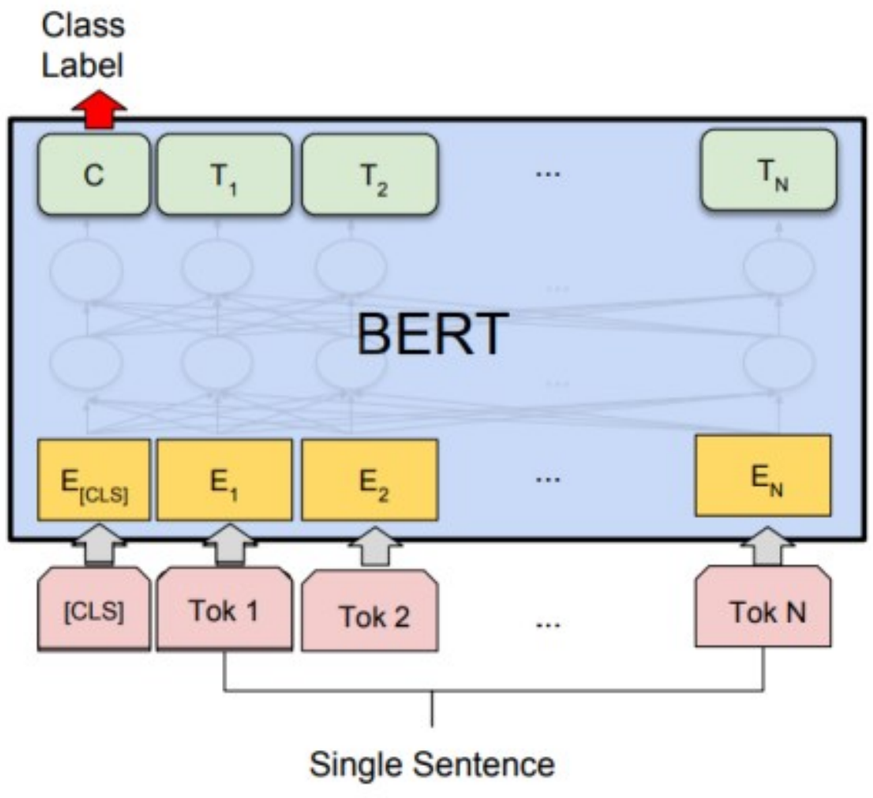
و هنا يتم وضع جملتين و يتم إدخالهم بهذه الطريقة حتي يتمكن بيرت من تحديد مدي توافقهم معا , مع التأكيد علي أن العنصر [SEP] يتم وضعه للفصل بين الجملتين



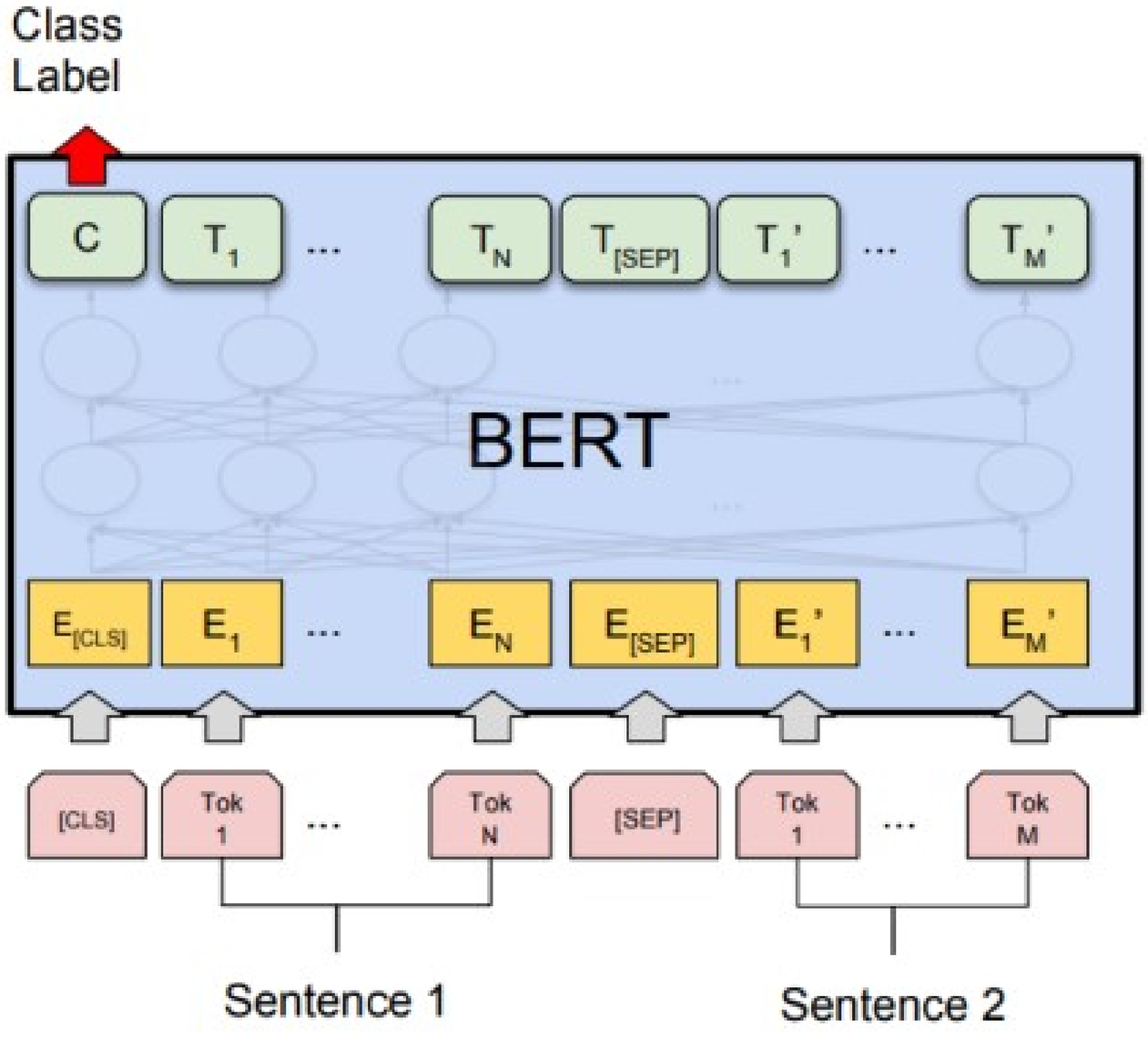
و هذه مقارنة بين تصميم بيرت و النماذج المختلفة مثل (OpenAI , ELMo(embedding from language model



و هنا تظهر الطبقة الاخيرة الخضراء التي نقوم بتدريبها , بينما الطبقات السابقة تكون مدربة بالفعل



و يكون هذا النموذج المشابهة لاستخدام جمل كاملة و ليس جملة واحدة ,حيث يحتوي علي [SEP]



كما أن بيرت قادر علي الاجابة عن سؤال محدد , مع اعطاءه قطعة نصية و يكون فيها الاجابة , ويكون قادر علي تحديد مكان الاجابة

- Input Question:

```
Where do water droplets collide with ice
crystals to form precipitation?
```

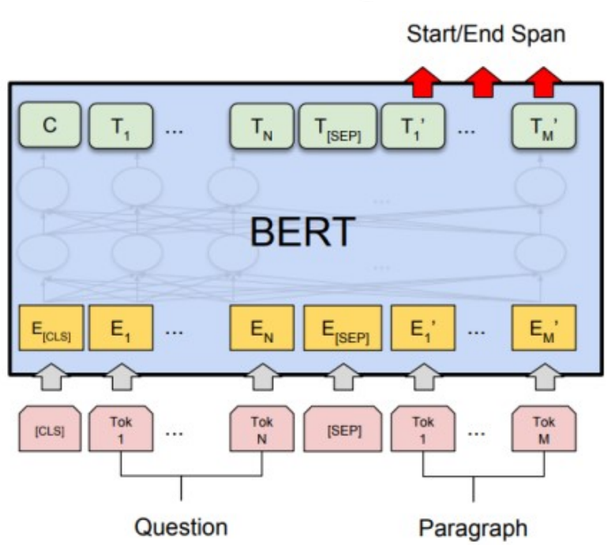
- Input Paragraph:

```
...  Precipitation forms as smaller droplets
coalesce via collision with other rain drops
or ice crystals within a cloud.  ...
```

- Output Answer:

```
within a cloud
```

و يكون تصميمها هكذا , حيث يتمكن بيرت من تحديد موضع الاجابة



_