التَّشابُهُ الدِّلاليُّ بينَ الجملِ العربيَّةِ من خلالِ تِقْنِيَة (BERT بيرت) الإلكترونيَّة: دراسة تقييميَّة حاسوبيَّة

Semantic Similarity between Arabic Sentences through BERT Electronic Technology: A Computational Evaluation Study

محمد مجدي لبيب*

Mohamed.labib@must.edu.eg

الملخص

يهدف البحث إلى تقييم وتقويم أداة قياس تشابه الجمل (Sentence similarity) الملحقة بـ(BERT) المعدة من (Google)، التي يعتمد عليها بشكل كبير في البحوث المهتمة بمعالجة اللُّغات الطَّبيعيَّة، خاصة في تحسين مخرجات التَّرجمة الآليَّة[1]، وذلك من خلال تتبع دقة مخرجاتها ودراسة تلك المخرجات، ثم ترجمة النتائج إلى إحصاءات توضح مدى دقة تعامل هذه الأداة المهمة مع اللُّغة.

وللوصول للهدف المنشود من البحث، تم الاعتماد على مدونة متوازية بين اللغة العربية واللغة الإنجليزية، ثم ترجمة عينة عشوائية من المدونة باللغة الإنجليزية على (Google Translate)[2]، ثم محاذاة نتائج الترجمة مع المدونة باللغة العربية، ثم إدخال أزواج الجمل المتحاذية (Patterns) باللغة العربية إلى (BERT)؛ لقياس مدى التشابه الدلالي بينها من خلال الأداة (Sentence similarity).

وأمكن البحثُ من خلال التطبيق العملي وتحليل مخرجات (BERT)، التوصل إلى مواضع الخلل التي تعيق عمل الأداة مع اللغة العربية، مقارنًا تلك

* مدرس الدراسات اللُّغويَّة – كليَّة اللُّغات والتَّرجمة – جامعة مصر للعلوم والتكنولوجيا.

النتائج بتعامل الأداة نفسها مع اللغة الإنجليزية، وكانت النتيجة في صالح اللغة الإنجليزية؛ حيث بلغت نسبة كفاءة الأداة معها حوالي (65%)، في مقابل (40%) مع اللُّغة العربيَّة.

وقد وضع البحث مقترحًا يسهم في تحسين مخرجات تعامل (BERT) مع اللغة العربية، مستندًا في ذلك على نتائج تحليل عينة الدراسة، والوقوف على أبرز الأخطاء التي لم تستطع الأداة تخطيها، مما قلل من كفاءتها.

الكلمات المفتاحية: BERT – التَّشابه الدِّلاليُّ بين النُّصوص – معالجة اللُّغات الطَّبيعيَّة – المدوَّنات اللغويَّة – التَّرجمة الآليَّة.

Abstract

This research aims to evaluate and calibrate the sentence similarity measurement tool associated with BERT (developed by Google), which is heavily relied upon in research concerning natural language processing, particularly in enhancing the outputs of machine translation. This is achieved by tracking the accuracy of its outputs and studying these outputs, then translating the results into statistics that illustrate the accuracy of this essential tool's handling of the language.

To achieve the desired research objective, a parallel corpus between Arabic and English was used. A random sample from the corpus in English was translated using Google Translate, followed by aligning the translation results with the Arabic corpus. Then, pairs of aligned sentences (Patterns) in Arabic were fed into BERT to measure the semantic similarity between them through the Sentence similarity tool.

Through practical application and analysis of BERT's outputs, the research identified the shortcomings that hinder the tool's performance with the Arabic language, comparing these results with the tool's performance in English. The outcome was in favor of the English language, where the tool's efficiency with it was about 65%, compared to 40% with Arabic.

The research proposed a solution that contributes to improving BERT's outputs with the Arabic language. This was based on the results of analyzing the study sample and identifying the most significant errors that the tool could not overcome, which reduced its efficiency.

Keywords: BERT, Semantic Text Similarity, Natural Language Processing, Linguistic Corpora, Machine Translation.

مقدمة

تقدم شركة جوجل (Google) عددًا من أدوات معالجة اللُّغات الطَّبيعيَّة، ومن أبرزها مجموعة أدوات (BERT) التي توفر واحدة من الأدوات المهمة التي تعمل على قياس نسبة التَّشابه الدِّلالي بين النُّصوص وهي (Sentence similarity)، بيد أن معالجة هذه التقنية للغة العربيَّة لم تصل دقتها إلى درجات الكفاءة المطلوبة، على عكس درجة كفاءتها عند معالجة اللُّغة الإنجليزيَّة، لذا فإن البحث، حاول دراسة هذه الأداة بشكل أكثر تفصيلًا وتطبيقًا؛ للوقوف على أسباب عدم دقتها عند معالجة اللُّغة العربيَّة، ومن ثَمَّ تقديم حلول لها.

التعريف بالأداة (BERT)

هي Bidirectional Encoder Representations from Transformers، المقدمة من جوجل (Google)، وهي عبارة عن تقنية حديثة من أساليب ومعالجات اللُّغات الطَّبيعيَّة، من تقنيات الشَّبكات العصبيَّة التي تمَّ اختراعها بهدف مساعدة الآلة على فَهم اللُّغة بطريقةٍ أقرب لفَهم البشر، ومن المهام التي تؤديها قياس نسبة التَّشابه بين النُّصوص.[3]

ومن أبرز المهام التي تقدمها توفيرُ إمكانيَّة قياس التَّشابه الدِّلاليّ بين الجمل من خلال خاصية (Sentence similarity)، والتي تلعب دورًا مهمًّا في بحوث معالجة اللُّغات الطَّبيعيَّة خاصة تلك المهتمَّة بالتَّرجمة الآليَّة، وصناعة المدوَّنات اللُّغويَّة المتوازية، والتي تعدُّ مصدرًا مهمًّا يتضمن الاختلافات بين النُّصوص الأصليَّة وترجماتها وبالتَّالي معرفة كيفيَّة نقل فكرة أو تعبير عبر لغتين أو أكثر؛ إذ توظَّف في مقارنة خصائص النَّصّ الأصليّ والنَّصّ المترجَم، كما أنَّها مصدر لُغويّ مهم لتدريب أنظمة التَّرجمة الآليَّة، وهي كذلك تعدُّ مخزنًا حاسوبيًّا

للتَّعبيرات الثَّابتة والتَّراكيب الاصطلاحيَّة أو المشتركة بين اللُّغة المصدر واللُّغة الهدف، كما أنَّها تساعد على اختيار الترجمات الأفضل اعتمادًا على السِّياق، وهي أيضًا تكشف الطَّبيعة المميزة للنُّصوص المتخصِّصة المختلفة، والتَّعرُّف على السِّمات الدِّلاليَّة والنَّحويَّة والتَّركيبيَّة بنصوص كلِّ لغة داخل المدوَّنة[4]؛ حيث إن الأداة تعطي كلَّ زوج من الجمل سواء في اللُّغة العربيَّة أو الإنجليزيَّة نسبة تقارب من خلال قياس درجة التَّشابه الدِّلاليّ بينهما؛ حيث تقع تلك النِّسبة بين $(0{,}0 \leq S.S \leq 1)$.

مادة الدراسة

اعتمد البحثُ في سبيل الوصول إلى هدفه، والتَّمكُّن من قياس كفاءة تعامل (BERT) مع اللُّغةِ العربيَّة، على مدوَّنة متوازية باللُّغتين العربيَّة والإنجليزيَّة، مكوَّنة من مليون زوج من اللُّغتين، ثم اختيار عيِّنة عشوائيَّة من المدوَّنتين قوامها (100,000) مائة ألف زوج من الجمل باللُّغتين، ثم ترجمة الجمل باللُّغة الإنجليزيَّة باستخدام الأداة (Google Translate)، ثم بناءُ مدونة متوازية تضم الجملَ الأصلية ونظيرتها من الجمل المترجَمة.

وتمَّ تقسيم عيِّنة الدِّراسة وَفْق درجة التَّشابه الدِّلاليّ بين زوجي الجملة العربيَّة إلى إحدى عشرةَ مجموعة / رتبة تمثِّل نسب درجات الَّتشابه التي يضعها (BERT)، يوضِّحها الجدول التَّالي:

جدول (1): تقسيم جمل مدوَّنة الدِّراسة بناء على نسبة التَّشابه بين زوجي جمل اللُّغة العربيَّة

Semantic similarity	Rank / الرتبة
1	[1]
0,9 ≤ S.S < 1	[2]
0,8 ≤ S.S < 0,9	[3]
0,7 ≤ S.S < 0,8	[4]
0,6 ≤ S.S < 0,7	[5]
0,5 ≤ S.S < 0,6	[6]
0,4 ≤ S.S < 0,5	[7]
0,3 ≤ S.S < 0,4	[8]
0,2 ≤ S.S < 0,3	[9]
0,1 ≤ S.S < 0,2	[10]
0,0 ≤ S.S < 0,1	[11]

طبيعة عمل (BERT)

يوفِّر (BERT) حزمةً من أدوات معالجة اللُّغات الطَّبيعيَّة على مختلف المستويات[5]، ومن الأدوات التي يوفِّرها وظيفة التَّشابه الدِّلالي (Semantic similarity)[6] التي من خلالها يمكن قياسُ مدى التَّشابه بين زوجين من الجمل، من حيث الدِّلالة والشَّكل.

ويعطي كلَّ زوج من الجمل أربعة أرقام بين كل رتبة والأخرى؛ حيث إنَّ النِّسب بين الرتبتين (0,9 ≤ S.S < 1) تتضمن التَّرقيم من (0,9999) و(0,9998) حتى التَّرقيم (0,9001) و(0,9000)، وهذا التَّغيُّر في النِّسب يدلُّ على وجود فارق بين الجملتين لكن بنسبة أو بدرجة معيّنة، مثال على ذلك، الجدول التَّالي:

M	Sentence	Human translation	Google translation	BERT
1	What should we do?	ماذا علينا أن نفعل؟	ماذا علينا ان نفعل؟	0.9999
2	you're lucky.	أنت محظوظ.	انت محظوظ.	0.9999

يلاحظُ من الجدول السَّابق أنَّ النِّسبة قلَّت درجة من (1) إلى (0,9999) بسبب وجود حرف مختلف بين الجملتين، ففي الجملة الأولى نجد أن الاختلاف بين همزة ألف المفردة (أن)؛ حيث كُتبت في ترجمة جوجل بدون همزة (ان)؛ ما تسبب في انخفاض درجة التَّشابه بين الجملتين، رغم أنَّهما متطابقتان في الدِّلالة. الأمر نفسُه تكرر في الجملة الثَّانية؛ حيث كتبت المفردة (أنت) في ترجمة جوجل بدون همزة (انت)، في حين أنَّها في التَّرجمة البشريَّة كتبت بهمزة قطع (أنت)، وهو ما تسبب في قلة نسبة التَّشابه بين الجملتين رغم تطابقهما في المعنى.

الأمر الذي لن يكون موجودًا إذا تمَّت مراجعة الجمل مراجعة لُغويَّة؛ إملائيَّة ونحويَّة، وتصحيحها في ضوء القواعد المعياريَّة للُّغة العربيَّة.

إن وظيفة التَّشابهِ الدِّلاليّ (Semantic similarity) الملحقة بـ(BERT) دورها قياس مدى التَّشابه بين زوجين من الجمل من حيث الدِّلالة، ولكن بتحليل مخرجات الأداة، لوحظ أن مجرد الاختلاف في شكل من أشكال الجملتين يؤدي إما إلى زيادة نسبة التَّشابه أو قلَّة نسبة التَّشابه، رغم أن الجملتين متطابقتان دلاليًّا بالفعل، وهو ما يعكسه الجدول التَّالي:

جدول (2): اختلاف نسبة التَّشابه بين الجمل على (BERT) بسبب وجود اختلاف في الشَّكل

M	Sentence	Human translation	Google translation	BERT
1	But why isn't she waking up?	ولكن لماذا لا تستيقظ؟	لكن لماذا لا تستيقظ؟	0.9998
2	I will be back.	سأعود.	سوف أعود.	0.9997
3	I will be back	سأعود	سوف أعود	0.9996
4	She's so lucky	إنها محظوظة جدا	هى محظوظة جدا	0.9995
5	I'm kidding.	أنا أمزح	أنا أمزح.	0.9994
6	In a moment In a moment	لحظة لحظة	في لحظة	0.9097
7	I love you, Scott...	أنا أحبك يا (سكوت)...	أحبك يا سكوت...	0.9096
8	That's wonderful then	هذا جيد	هذا رائع إذن	0.9060
9	It's on fire... and it's green...	إنها في النار ...وهي خضراء	إنها مشتعلة ... وهي خضراء...	0.9045
10	No, sir. It was perfect.	لا يا سيدي هذا مثالي	لا سيدي. كان مثاليا.	0.9028
11	Go away from here	اخرجي من هنا	اذهب من هنا	0.9020
12	You can hold me if you're scared.	يمكنك الإمساك بي إن كنت خائفا.	يمكنك أن تمسكني إذا كنت خائفا.	0.9013
13	What are you saying?	ماذا قلت؟	ماذا تقول؟	0.9012
14	You can name the price.	حدد أنت السعر.	يمكنك تسمية السعر.	0.9007
15	you can forgive me?	تسامحني؟	يمكنك أن يغفر لي؟	0.9004
16	I know that	أعرف أنك حقا تحبينه	أعلم أنك أحببته حقا.	0.9003

	you really truly liked him.			
17	Didn't I tell you?	أنا لم أخبرك بعد؟	ألم أقل لك؟	0.9001
18	There he is now.	هاهو الآن	ها هو الآن	0.9978

الجدول (2) يستعرض عددًا من أزواج الجمل التي تنتمي إلى الرتبة الثَّانية وفق مخرجات التِّشابهِ الدِّلالي على (BERT)، ويرجِعُ تفاوتَ نسب التَّشابه لعدة أسباب، منها زيادة أو نقصان كلمة في جملة مقابل الجملة الأخرى، بالإضافة إلى اختلاف الكلمات الموظَّفة، مثل زمن الفعل مضارع أو ماضٍ، كما في المثال (13)، والاختلاف بين المصدر والفعل كما في المثال (12)، ولكن هذه الاختلافات لم تؤثِّر على رتبة التَّشابه بين الجملتين؛ حيث إن جميعها كانت في الرتبة الثَّانية في درجة التَّشابه.

وكلما زادت الفوارق بين الجملتين قلَّت نسبة التَّشابه بينهما؛ حتى تصل إلى النِّسبة بين $(0{,}0 \leq S.S < 0{,}1)$، مثل:

M	Sentence	Human translation	Google translation	BERT
1	She still looks perfectly healthy.	تدهورت صحتها.	لا تزال تبدو بصحة جيدة.	0.0769
2	I'm the one who provoked him first.	أنا الذي أثرته أولا	0	0.0248

وفق الجدول السَّابق فإنَّه كان من المتوقَّع أن تكون درجة التَّشابه بين أزواج الجملتين تساوي صفرًا (0)؛ نظرًا لعدم وجود تشابه إطلاقًا بين زوجي الجملة الأولى أو الثَّانية، ورغم ذلك فإن BERT أعطى لهما نسبة تشابه على قلتها، فحتى الجملة التي ليس لها زوج مقابل لها (الجملة رقم 2) أعطى لها درجة

تشابه (0,0248)، حتى وإن كانت في لغة الرياضيات ليس لها دلالة لكنها هنا تحمل دلالة وجود نسبة تشابه حتى وإن كانت ضئيلة.

ويتضح من العرض السَّريع السَّابق أن رتبةَ التَّشابه ودرجتَه تتوقف حتى على علامات التَّرقيم والمسافات بين الكلمات، مما يخلق اختلافًا بين الجمل وهو ما يؤثِّر على دقَّة نتائج الأداة، وربما كان مآل هذا الاختلاف هو الاختفاء إذا تمَّت مراجعةُ الجمل وتدقيقُها قبل إدخالها إلى الأداة المستخدمة.

استخدام (BERT) في التَّطبيقات اللُّغويَّة:

1) من أبرز استخدامات (BERT) توظيفه في بناء المدوَّنات اللُّغويَّة المتوازية المتحاذية المستخدمة في تحسين مخرجات التَّرجمة الآليَّة، فمن خلاله يمكن اختيار النَّماذج التَّطبيقيَّة، التي من خلالها يتمُّ تقييم دقَّة مخرجات التَّرجمة الآليَّة وكفاءتها، والوقوف على مسببات ضعف أدائها ووضع منهجيَّة لحلها وبالتَّالي تحسين المخرج النِّهائي؛ حيث يمكن من خلاله استخلاص العيِّنة اللازمة للتَّطبيق، فبدلًا من دراسة نصوص المدوَّنة كاملة، فإنَّ (BERT) يساعد في اختيار الجمل التي تتضمن إشكاليَّات تعيق ترجمتها بصورة صحيحة.

2) مراجعة نصوص المدوَّنات الخام والتَّأكُّد من صلاحيتها لأغراض الدراسات المختلفة؛ حيث إن (BERT) يصنِّف الجمل وَفْق درجة تشابهها الدِّلاليّ، ومختلف المستويات اللُّغويَّة، بالإضافة إلى الشَّكل الكتابي للُّغة.

خلل في مخرجات (BERT) حول اللُّغة العربيَّة

بالنَّظر في مخرجات (BERT) حول مدوَّنة اللُّغة العربيَّة، لوحظ وجود بعض التَّصنيفات التي يشوبها الخلل؛ إذ إنَّ الأداة أعطت زوجًا من الجمل نسبة تشابه معيَّنة لسبب ما، ولكن بالنَّظر في باقي الأزواج، وُجد أن الخطأ نفسه

تكرر مع زوج آخر من الجمل، لكن الأداة أعطته نسبة تشابه أكبر أو أصغر من مثيله، كما في الأمثلة التَّالية:

جدول (3): خلل في (BERT) يكرر الخطأ نفسه بنسب تشابه مختلفة بين أزواج الجمل

M	Sentence	Human translation	Google translation	BERT
1	You're so cute.	أنت لطيف جدا.	انت لطيف جدا.	0.9998
2	I am ill	أنا مريض	انا مريض	0.9996
3	I don't like you.	أنا لا أحبك.	انا لا احبك.	0.9870
4	Where are my friends?	أين أصدقائي؟	اين اصدقائي؟	0.9960
5	It's the sun.	إنها الشمس	إنها الشمس.	0.9971
6	They're in there.	إنهم هناك	إنهم هناك.	0.9970
7	ln the dark.	في الظلام	في الظلام.	0.9968
8	I'm behind you.	أنا خلفك	أنا خلفك.	0.9964
9	I will see him now.	سأراه الآن	سأراه الآن.	0.9955
10	Never.	أبدا	أبدا.	0.9886
11	There.	هناك	هناك.	0.9895
12	You have to stand.	عليك أن تقف	عليك أن تقف.	0.9887
13	Full stop.	نقطة	نقطة.	0.9878
14	It is sturdy.	إنه قوي	إنه قوي.	0.9873
15	Traitor	الخائن	خائن	0.9861
16	My hesitation?	ترددي ؟	ترددي؟	0.9894
17	Why are you crying?	لماذا تبكين؟	لماذا تبكين	0.9850
18	Just, whatever happened	فقط مهما حدث	فقط، مهما حدث	0.9847
19	I haven't slept very well.	لم أنام جيدا	لم أنم جيدا.	0.9851
20	Don't come to me anymore.	لا تأتي إلي بعد الآن	لا تأتي إليّ بعد الآن	0.9845
21	There he is now	هاهو الآن	ها هو الآن	0.9978

يظهر الجدولُ خللًا في منهجية (BERT) لتصنيف أزواج الجمل وَفْق نسبة التَّشابه الدِّلالي بينها؛ حيث هناك مجموعة من الجمل ورد تباينٌ بينها فيما يتعلَّق بنسبة التَّشابه رغم أن الاختلافَ واحدٌ بينها، على سبيل المثال الجملة رقم (3) كتبت مفرداتها في التَّرجمة الآليَّة بدون همزة (انا لا احبك) وفي المقابل كانت ترجمتها البشريَّة تنصُّ على وجود الهمزة (أنا لا أحبك)، ما جعل (BERT) يسجِّل لها نسبة تشابه (0.9870)، بينما الجملة رقم (4) ورد فيها الاختلاف نفسه بين التَّرجمة البشريَّة والتَّرجمة الآليَّة لكنه منحها نسبة تشابه مختلفة هي (0.9960).

أيضًا فإنَّ (BERT) يضع في الاعتبار الاختلافَ بين الجملتين في علامات التَّرقيم، ففي المثال (7) الجملتان متطابقتان في الكتابة وترتيب الكلمات وبدون أخطاء إملائيَّة لكنَّه لم يعطها نسبة تشابه كاملة، بل كانت نسبة التَّشابه (0.9968)؛ نظرًا لوجود علامة الترقيم (.) في ترجمة جوجل وعدم وجودها في التَّرجمة البشريَّة، وهو ما تكرر في الأمثلة رقم (5)، و(6)، و(8)، إلى المثال رقم (14).

وهناك اختلاف آخر لم يراعه (BERT) عند وضع نسب التَّشابه لأزواج الجمل، وهو وجود مسافة قبل علامات التَّرقيم، كما في المثال (16)، أو وجود علامات التَّرقيم من عدمها، كما في المثال رقم (17)، و(18)؛ حيث إن الجمل تستحق نسبة تشابه (1)، ولكن بسبب هذه الاختلافات البسيطة التي تعود إلى أخطاء الكتابة تم إعطاؤها نسب تشابه متفاوتة وأقل من الصحيحة، والمثال رقم (20)، لم تعد الآلة الجملتين متتطابقتين تمامًا؛ نظرًا لوجود التصاق بين كلمتين، وعدم التصاقهما في الجملة الموازية لها، في (هاهو)، و(ها هو).

الأخطاء النَّحويَّة أيضًا من المؤثرات بشكل كبير في كفاءة تقدير (BERT) لنسبة التَّشابه بين أزواج الجمل، كما في المثال رقم (19)، في الجدول نفسه؛

حيث إنَّ الفعل (أنام) أتى في التَّرجمة الآليَّة مجزومًا بـ(لم)، فيما لم يتم جزمه في التَّرجمة البشريَّة رغم أنَّه مسبوق بأداة الجزم (لم)، ما أدى إلى أن تكون نسبة التَّشابه بين الجملتين (0.9851)، فيما إذا تمت مراجعة الجملتين وتصحيح ما بهما من خطأ، فإنَّ درجة التشابه بينهما ستكون (1).

أما المثال رقم (20)، فيعكس تأثير علامات الضبط على دقة نتائج (BERT)، فإذا كانت الجملتان متطابقتين تمامًا إلا في علامة ضبط واحدة، فإن نسبة التشابه تقل.

كل هذا يؤكد أن مخرجات (BERT) ليست دقيقة نسبيًّا، وأن مادة المدوَّنة الخام يجب أن تخضع لعمليَّات المراجعة اللُّغويَّة والتدقيق؛ لضمان دقة النتائج، لكن هذا لا يقلل من أهمية الأداة.

وأيضًا من العوامل المؤثِّرة على عمل (BERT) في تحديد نسبة التَّشابه بين أزواج الجمل، ترتيب الجمل؛ حيث إن الجمل قد يكون لها نفس المعنى والدِّلالة، لكنه أعطاها نسبة تشابه أقل من التَّطابق بسبب اختلاف ترتيب الكلمات كما في المثال رقم (1) في الجدول التالي، وأيضًا استخدام مفردات مختلفة لكنها تفيد الدِّلالة نفسها مثل (أمك)، و(والدتك)، المفردتان تؤديان المعنى نفسه في الجملتين (لقد فقد أمه)، و(لقد فقد والدته)، ورغم تطابق الجملتين في الدِّلالة بيد أن (BERT) اعتبر أنَّ التَّشابه بينهما أقل من حد التطابق، فأعطى لهما نسبة تشابه بلغت (0.9818)، وهو ما يظهر في المثال (2)، من الجدول نفسه.

M	Sentence	Human translation	Google translation	BERT
1	The rhythm of production will be affected.	إيقاع الإنتاج سيتأثر	سيتأثر إيقاع الإنتاج	0.9850
2	He lost his mom.	لقد فقد أمه	لقد فقد والدته	0.9818

ومن أشكال الخلل التي ظهرت في مخرجات (BERT) أنَّه قد يعالج الجملة نفسها مرتين ولكن يعطي في كل مرة نسبة تشابه مختلفة؛ بسبب وجود اختلاف بين تركيب الجملتين مهما صغر، كما في الجدول التالي:

جدول (4): يشرح كيف يعطي (BERT) زوجًا متشابهًا من الجمل نسبة تشابه مختلفة

M	Sentence	Human translation	Google translation	BERT
1	How embarrassing...	يا للإحراج	كم هذا محرج...	0.7982
	How embarrassing.	يا للإحراج	كم هذا محرج.	0.7841
2	How do you like Japan?	كيف وجدتم اليابان ؟	كيف تحب اليابان؟	0.7879
	How do you like Japan?	كيف وجدتم (اليابان) ؟	كيف تحب اليابان؟	0.7718
3	Move aside	إبتعدوا	تحرك جانبا	0.7621
	Move aside!	ابتعدوا.	تحرك جانبا!	0.7573
4	Caught you	سأنال منك	امسكتك	0.7463
		سأنال منك	امسكتك	0.7323
5	he is gentle, and kind.	إنه لطيف وطيب	هو لطيفة ولطيفة.	0.7586
	She is gentle, and kind.	إنه لطيف وطيب !	هي لطيفة ولطيفة.	0.7267
6	I wanted it than they want	أود أكثر بكثير مما خططوا إلي.	أردت ذلك مما يريدون	0.6996
	I wanted it than they want	أود أكثر بكثير مما خططوا إلي.	أردت ذلك مما يريدون.	0.6956
7	I will finish off Zero...!	سأقتل زيرو	سأنهي الصفر ...!	0.6877
	I will finish off Zero...	سأقتل زيرو	سوف أنهي الصفر ...	0.6174
8	he always goesget lost	دائما يردد اغرب	هو دائما يذهب ، ويضيع	0.7477
	he always goes "get lost"	دائما يردد "أغرب".	هو دائما يذهب "يضيع"	0.6855
9	One road was closed, others open.	يغلق باب يفتح غيره.	تم إغلاق طريق واحد، وفتح البعض الآخر.	0.6762
	One road was closed, others open.	يغلق باب، يفتح غيره	تم إغلاق طريق واحد، وفتح البعض الآخر.	0.6582
10	Just what I wanted	كنت آمل ذلك	فقط ما أردت	0.6443
	Just what I wanted!	كنت آمل ذلك	فقط ما أردت!	0.6188
11	I will finish off Zero...!	سأقتل زيرو	سأنهي الصفر ...!	0.6877
	I will finish off Zero...	سأقتل زيرو	سوف أنهي الصفر ...	0.6174
12	This is your punishment!	هذه هي العدالة	هذا عقابك!	0.7183
	This is your punishment	هذه هي العداالة	هذا عقابك	0.6026

يلاحظ أن الجملة (How embarrassing) تكررت في المدوَّنة مرتين، وأظهرها (BERT) مرتين، لكن في كل مرة وضع لها نسبة تشابه مختلفة، وهذا يعود إلى وجود علامة ترقيم مختلفة بين الجملتين، فالأولى تضمنت ثلاث نقاط (...)، والثَّانية تضمنت نقطة واحدة (.)، ورغم وقوع الجملتين في الرتبة نفسها، لكن درجة التَّشابه بين أزواج الجمل في المرتين مختلفة. حتى إنَّ وجود مسافة في جملة من الجملتين يؤدي إلى تصنيف الزوجين في نسب تشابه مختلفة، كما يظهر من المثال (4)، بسبب وجود مسافة قبل الجملة (سأنال منك)، منحها (BERT) نسبة تشابه مع الجملة المكافئة لها بلغت (0.7323)، في حين أن نسبة التشابه كانت (0.7463) عند اختفاء هذه المسافة.

هذا العرض يؤكد أنَّه إذا تمت مراجعة المدوَّنة الخام وتصحيح ما بها من أخطاء فإن عدد الجمل المتطابقة في الدِّلالة سترتفع.

وعند فحص الرتب الأقل تشابهًا بين أزواج الجمل بداية من الرتبة الخامسة، فإن (BERT) يعطي نسب تشابه بين أزواج الجمل، رغم أن تقييمهم بشريًّا يؤكد أن لا تشابه بينهما، كما في الجدول التَّالي:

جدول (5): وجود خلل في نسبة التشابه بين أزواج من الجمل ليس بينهما تشابه

M	Sentence	Human translation	Google translation	BERT
1	The Japs were sharp.	كانت Japs حادة.	اليابانيون كانوا حادين.	0.5997
2	Why are these books doing in my house?	لماذا هذه الكتب تفعل في منزلي؟	لم هذا في منزلي؟	0.5991
3	there's no way I'll tell you that	لا توجد طريقة سأخبرك بها	لن أخبرك أيـها الخائن	0.5985
4	someone there?	شخص ما هناك؟	ثمة أمر.. هنا؟	0.5983
5	It's not bad.	ليس سيئا.	تبدو جيدة	0.5982
6	You're not going anywhere.	كنت لا أذهب إلى أي مكان.	أنت لن تذهبين إلى أي مكان	0.5980
7	But this is unheard of.	لكن هذا لم يسمع به.	ولكنني لم أسمع بذلك من قبل	0.5978
8	Extraordinary.	نادر.	مذهل	0.5977
9	Somewhere, in his soul, he was a Pengeran.	في مكان ما، في روحه، كان Pengeran.	هنالك أمير في مكان ما في أعماق روحه	0.5977
10	If you need anything, you can reach me on this intercom.	إذا كنت بحاجة إلى أي شيء، يمكنك الوصول إلي على هذا الاتصال الداخلي.	إن احتجتم لأي شي اتصلوا علي بواسطة الأنتركم.	0.5976
11	Why'd they ask you to do that?	لماذا طلبوا منك أن تفعل ذلك؟	لماذا يقولون لك أن ترتدي هذا ؟	0.5976
12	I feel myself to appear in front of him.	أشعر بنفسي لأظهر أمامه.	أخدع نفسي وأنتظرها لتعبره	0.5974
13	Submit your daughter to me, and I'll set you free.	أرسل ابنتك لي، وسأطلق سراحك.	وافق على زواجي من ابنتك فحسب وسأحررك أبدا	0.5971

Score: 0.5038	جناح؟	غرفة جناح؟
Score: 0.5042	الذي يطير فوق ال..	الطائر الأزرق الذي يطير بعيدا.
Score: 0.5044	هذه حقًا مزحة غير مضحكة	أجد نكاتك مسيئة للغاية.
Score: 0.5045	كيف تورط قائدالفريق في هذ الوضع؟	كيف وقع الزعيم كانغ في هذا الموقف؟
Score: 0.5047	على الأقل أمكنني رؤيتك	هذه هي المرة الأخيرة بالنسبة لي لرؤيتك.
Score: 0.5047	تحلى بالقوة	يجب أن تبتهج.
Score: 0.5047	حقًا، إنه طائر	جدي، انه عصفور!
Score: 0.5051	أين هو زيرو الآن	أين هيك هو الصفر؟
Score: 0.5051	هل خرجت للمساعده ؟	سيكون الأمر غريبا إذا لم يفعل.
Score: 0.5051	إنها ليست مثلنا في شيء.	هو لا يحبنا.
Score: 0.5057	لذا فقد خذلك العالم	وسأكون في النهاية الشخص الذي يسيطر على العالم.
Score: 0.5057	تعنين من المسابقة.	تقصد العرض؟
Score: 0.5058	مساء الخيـر	ما تريده هو درعك الأبيض.
Score: 0.5061	أرجوكِ لا تتكلمي لقد ذهب زيرو إلى كامين جيما	إخفاء حقيقة أنني مصاب ...
Score: 0.5061	أرجوك أخبريني ما الذي قالته.	من فضلك قل لي ماذا تقول.
Score: 0.5063	مؤلم	مرارة - مر.
Score: 0.5072	من سقطت هنا؟	هذا حيث كان الحذاء؟
Score: 0.5074	هل يلتمس إحسانا؟	هل يطلب معروفا؟
Score: 0.5078	الأميرة.	أميرة...
Score: 0.5081	هل هذه.	يا بلادي؟

وجود خلل في توافق نسبة التشابه بين أزواج الجمل

يوضِّح الشَّكل السَّابق وجود خلل ما في حساب (BERT) لنسبة التَّشابه بين أزواج جمل ليس بينهما علاقة أو لا يصلح التزواج بينهما؛ على سبيل المثال فإن الجملتين (هل هذه) و(يا بلادي)، ليس بينهما علاقة تشابه لكن الأداة اعتبرت أنهما متشابهتان بنسبة (0,5081)، وهو غير صحيح.

مع الأخذ في الاعتبار أن هذه الفئة أو الرتبة الخامسة قد تتضمن مجموعة من الجمل التي قد ترى التِّقنيَّة وجه تشابه بين زوج من الجمل على عكس الخبرة البشريَّة التي ترى أن لا تشابه بينهما، مثل الجملة (تحلى بالقوة) في مقابل الجملة (يجب أن تبتهج)؛ حيث كانت نسبة التَّشابه بينهما (0,5047)، فربما قاربت الآلة بينهما على أنهما تعبران عن نصيحة، رغم أن ذلك لا يتوافر في جمل أخرى تفوقهما في نسبة التشابه مثل (من سقط هنا؟) والجملة الموازية لها (هذا حيث كان الحذاء؟)؛ فرغم عدم التشابه بينهما، لكن (BERT) اعتبرهما متشابهتين بنسبة بلغت (0,5072) متخطية جملتي (تحلى بالقوة + يجب أن تبتهج).

وللوقوف على المشكلات التي تعيق عمل الأداة المستهدف تقييمها وتقويم عمليها، تم اختيار عيِّنة عشوائيَّة مكوَّنة من (100,000) مائة ألف زوج من اللُّغة العربيَّة، ثم عرضها على (BERT) لدراسة نسبة التَّشابه بين أزواج الجمل داخلها.

ثم تم اختيار عيِّنة طبقيَّة من المادة المختارة؛ حيث يمثِّل كلَّ رتبة (1067) ألفٌ وسبعة وستون زوجًا من الجمل تتضمن السِّمات والخصائص التي تغلب على جمل الرتبة المقصودة الرتب الإحدى عشرة.

وقد تم وضع مجموعة من الرُّموز للتَّعبير عن حالة كل زوج من أزواج الجمل الواردة في جداول الرتب الإحدى عشرة، توضيحًا للفكرة وتوفيرًا للوقت والجهد، كما أنَّها تيسِّر إجراء العمليَّات الإحصائيَّة على المدوَّنة.

ويوضِّح الجدول التَّالي الرُّموزَ المستخدمة؛ حيث كلّ رمز يعبِّر عن أسباب تقارب نسب التَّشابه بين أزواج الجمل، أو أسباب تباعدها، وهل هي متفقة في المعنى والمبنى، أو متفقة في المعنى دون المبنى، أو متفقة في المبنى دون المعنى، وهكذا.

جدول (6): رموز الدراسة في تصنيف أزواج الجمل وفق درجة التشابه بينها

م	الرَّمز	درجة التَّشابه	سبب التَّباعد
1	100	الجمل متشابهة تمامًا (معنى + مبنى)	
2	¥	الجمل متشابهة في الدِّلالة مختلفة في المبنى	وجود فارق بسيط مثل علامات التَّرقيم والمسافات
3	€	الجمل متشابهة في الدِّلالة مختلفة في المبنى	الأخطاء الإملائيَّة
4	☋	الجمل متقاربة في الدِّلالة	خلل في (BERT)
5	ϕ	الجمل غير متشابهة	ترحيل الجمل في المدوَّنة الخام
6	X	الجمل غير متشابهة	خلل في تركيب الجمل الإنجليزيَّة أدى إلى خلل في التَّرجمة الآليَّة
7	Ⓜ	نسبة التَّشابه قليلة	ترجمة الأعلام حرفيًّا
8	µ	الجمل غير متشابهة	وجود حروف لاتينيَّة
9	I	الجمل غير متشابهة	معنى مجازي

نتيجة اختبار العيّنة العشوائيَّة

وكانت نتيجة مسح عدد من الجمل التي تقع في نطاق كل رتبة، كما يلي:

1) الرتبة الأولى (SS = 1)

جدول (7): عينة من مخرجات (BERT) تنتمي للرتبة الأولى طبقًا لنسبة التشابه

M	Sentence	Human translation	Google translation	BERT
1	The weather stops here.	يتوقف الطقس هنا.	يتوقف الطقس هنا.	1
2	Pretty smart.	ذكي جدا.	ذكي جدا.	1
3	We walked on our knees until we got to the centre.	مشينا على ركبنا حتى وصلنا إلى المركز.	مشينا على ركبنا حتى وصلنا إلى المركز.	1
4	What have we got here?	ماذا لدينا هنا؟	ماذا لدينا هنا؟	1
5	Who killed Mother?	من قتل أمي؟	من قتل أمي؟	1
6	Good morning, Belle	صباح الخير يا بيل	صباح الخير يا بيل	1
7	Smell	رائحة	رائحة	1
8	Mrs wrong	السيدة مخطئة	السيدة مخطئة	1
9	No, sir Not me	لا يا سيدي ليس أنا	لا يا سيدي ليس أنا	1
10	you saved my life.	لقد أنقذت حياتي.	لقد أنقذت حياتي.	1

الملاحظات والاستنتاجات

1) كلُّ أزواج هذه الرُّتبة متطابقة تمامًا؛ نظرًا لاتفاقها في المعنى والمبنى.
2) تتأثَّر مخرجات (BERT) بوجود أي اختلاف بسيط بين زوجي الجملة، حتى لو في علامة ضبط أو علامة ترقيم، وللتَّأكُّد من ذلك فقد أجرى البحث تجربة على عيِّنة عشوائية من أزواج جمل المدوَّنة قبل المراجعة وبعدها، فوجد أن عدد الجمل المتطابقة ارتفع بعد التصحيح؛ لأن زوجي الجملة أصبحا متطابقين تمامًا.

مخرجات BERT بعد المراجعة اللغوية والتهيئة

Semantic similarity	عدد الجمل قبل التهيئة	عدد الجمل بعد التهيئة	عدد الجمل بعد التهيئة2	الفرق
1	524	957	1038	

2) الرتبة الثانية ($0{,}9 \leq S.S < 1$)

جدول (8): عينة من مخرجات (BERT) تنتمي للرتبة الثانية طبقًا لنسبة التشابه

M	Sentence	Human translation	Google translation	BERT
1	I'm sure of it.	أنا متأكدة من ذلك.	أنا متأكد من ذلك.	0.9998
2	I will be back.	سأعود.	سوف أعود.	0.9997
3	I don't know.	أنا لا أعلم	لا أعلم.	0.9984
4	Do you know where we are?	أتعرفين أين نحن؟	هل تعلم أين نحن؟	0.9884
5	Traitor	الخائن	خائن	0.9861
6	Until when?	إلى متى؟	حتى عندما؟	0.9198
7	He likes it very strong.	إنه يحبه بشدة	إنه يحبها بشدة.	0.9143
8	She ran off.	لقد هربت.	هربت.	0.9081
9	Bar that door	أغلق ذلك الباب	حظر هذا الباب	0.9063
10	Johnny doesn't know anything?	جوني لا يعرف شيئا	جوني لا يعرف شيئا؟	0.9001

الملاحظات والاستنتاجات

1. ترقى جمل هذه المجموعة في نسبة التَّشابه بين أزواج الجمل إلى مرتبة التَّطابق التَّام، ويمكن ضمُّها للمرتبة الأولى؛ إذ إنَّها تؤدي المعنى ذاته، لكن مع وجود اختلافات طفيفة، إما في الاستخدام أو في المبنى والتَّركيب، مثل فارق التَّذكير والتَّأنيث.
2. من الفوارق أيضًا بين أزواج الجمل في هذه المجموعة، استخدامات حروف المعاني أو الكلمات الوظيفية، مثل استخدام (السين) و(سوف)، و(هل) و(الهمزة).
3. استخدام المترادفات أيضًا مثل (تظن)، و(تعتقد).
4. اختلاف دلالات الضَّمائر؛ حيث في التَّرجمة اليدويَّة الضَّمير وُجِّه لجمع المخاطب، أما في التَّرجمة الآليَّة الضمير موجَّه للمفرد المخاطب، وهذا يتوقَّف على سياق الكلام. كما في المثال (8، 14).

الذِّكر والحذف أيضًا من شأنها أن تؤثِّر على مخرجات (BERT) وتجعله يقلل من نسبة التَّشابه بين الجملتين رغم أن الدِّلالة واحدة، كما في ذكر الضمائر في المثال (29، 30، 31).

والخلاصة في جمل هذه المجموعة أنها متقاربة إلى حد كبير فيما بينها على مستوى الدِّلالة والمبنى، وتصنيف (BERT) بنسب تشابه أقل من المجموعة الأولى يرجع إلى اختلافات دلاليَّة نابعة من الضمائر والاستخدام، أو اختلافات في المبنى.

3) الرتبة الثَّالثة ($0,8 \leq S.S < 0,9$)

جدول (9): عينة من مخرجات (BERT) تنتمي للرتبة الثالثة طبقًا لنسبة التشابه

M	Sentence	Human translation	Google translation	BERT
1	That's the most possible scenario.	إنه السيناريو المحتمل	هذا هو السيناريو الأكثر احتمالا.	0.8998
2	I mean, I don't mind.	أعني أنا لا أعارض	أعني، لا أمانع.	0.8997
3	I accepted his apology	لقد قبلت اعتذاره.	قبلت اعتذاره	0.8997
4	your hideout... is he the only one?	ملجؤك... .. هل هو الوحيد؟	مخبأك ... هل هو الوحيد؟	0.8991
5	He was stupid.	إنه أحمق	لقد كان غبيا.	0.8851
6	I've got it.	لقد فهمت.	لقد حصلت عليه.	0.8578
7	Life is very debilitating	الحياة مروعة جدا	الحياة منهكة للغاية	0.8699
8	Don't you work here?	ألا تعمل هنا؟	الا تعمل هنا	0.8457
9	I'm so lonely.	كنت أشعر بالوحدة	انا وحيد جدا.	0.8152
10	Finished	انتهى	تم الانتهاء منه	0.8034

الملاحظات والاستنتاجات

1) تضمنت هذه الرتبة مجموعة من الجمل غلب عليها تقارب التشابه الدِّلالي، مع اختلاف كبير في المبنى والتركيب، وهذا الاختلاف في المبنى إما أنه نتيجة استخدام المترادفات، أو التعبيرات التي تعطي الدلالة نفسها كما في (لا أمانع)، و(لا أعارض) في المثال رقم (2).

2) اختلاف في درجة التشابه بسبب زيادة في المبنى كما في ذكر (لقد) في المثال (3)، وأيضًا الدور الذي يلعبه وجود علامات الترقيم في المثال ذاته.

ورغم أن العدد الأكبر من أزواج جمل هذه المجموعة يعطي نسبة دلالة متقاربة، وأن سبب قلة نسبة التشابه يعود إلى التركيب واختلاف المبنى، بيد أنها قد تتضمن جملًا بها خللٌ في التركيب، أو ركاكة، مثل دلالة الأزمنة كما في المثال (5)، والسمات التداولية داخل الجمل كما في المثال (6).

مع الالتفات إلى أن الاختلاف قد يكون بسبب وجود خلل في الجملة الإنجليزيَّة، كتشابك الكلمات، أو حذف كلمة كالأفعال المساعدة، واستخدام الكلمات ذات الدلالة المتقاربة مثل (ملجأ) و(مخبأ) في المثال (4).

ويمكن أن تصنَّف مشكلات أزواج هذه الجمل وفق السِّمات التَّداولية.

4) الرُّتبة الرَّابعة ($0,7 \leq S.S < 0,8$)

جدول (10): عيِّنة من مخرجات (BERT) تنتمي للرتبة الرابعة وفْقْ نسبة التشابه

M	Sentence	Human translation	Google translation	BERT
1	Who would dare rob the betrothed of the great Khan?	من الذي يجرؤ على سرقة خطيبة خان المعظم؟	من يجرؤ على سرقة خطيب الخان العظيم؟	0.7893
2	You've been married to a boy and an old man.	لقد تزوجت من فتى ثم من رجل كهل	لقد تزوجت من ولد ورجل عجوز.	0.7996
3	And what's on the top of the water to you	وما هو الشيء الذي تراه فوق المياه	وماذا على سطح الماء لك	0.7889
4	on that mercenary ground, won't you	من منطلق الاستئجار ألا توافقين أن أستبد	على أرض المرتزقة ، ألا توافق على	0.7885

	agree to let me hector you a little?	بك قليلا؟	السماح لي بمراوغتك قليلاً؟	
5	he's always got his head in the clouds	هو دائما شارد الذهن هكذا	دائما ما يكون رأسه في الغيوم	0.7876
6	Take the clouds from your eyes.	أبعد الغشاوة عن عينيك.	خذ الغيوم من عينيك.	0.7735
7	Put yourself in my shoes	ضع نفسك مكاني	ضع نفسك في حذائي	0.7509
8	That wouldn't do you no good.	هذا ليس في مصلحتك	هذا لن يفيدك.	0.7703
9	Quit stalling.	لا تماطل.	توقف عن المماطلة.	0.7635
10	You've cheated.	أنت مخادع	لقد غشيت.	0.5859

الملاحظات والاستنتاجات

1) من المشكلات التي ظهرت في هذه المجموعة وأدت إلى تباعد نسبة التَّشابه بين أزواج الجمل عدم مراعاة دلالة المذكر والمؤنث.

Sentence	Human translation	Google translation
Who would dare rob the betrothed of the great Khan?	من الذي يجرؤ على سرقة خطيبة خان المعظم؟	من يجرؤ على سرقة خطيب الخان العظيم؟

2) استخدام التَّعبيرات الاصطلاحيَّة والمجازيَّة كما في:

Sentence	Human translation	Google translation
he's always got his head in the clouds	هو دائما شارد الذهن هكذا	دائما ما يكون رأسه في الغيوم
Take the clouds from your eyes.	أبعد الغشاوة عن عينيك.	خذ الغيوم من عينيك.
Take the clouds from your eyes and see me as I really am!	أبعد الغشاوة عن عينيك وشاهد حقيقتي	خذ الغيوم من عينيك وانظر لي كما أنا حقا!
Put yourself in my shoes	ضع نفسك مكاني	ضع نفسك في حذائي

3) ويلاحظ أيضًا أن هناك جملًا تعطي الدِّلالة نفسها، لكن (BERT) اعتبرها ذات نسبة تشابه قليلة بسبب فروقات طفيفة.

Sentence	Human translation	Google translation
I'd like to hear it.	أريد الاستماع اليه.	أود أن أسمعه.
It just keeps moving.	إنها تواصل الحركة	إنها فقط تستمر في التحرك.
That wouldn't do you no good.	هذا ليس في مصلحتك	هذا لن يفيدك.
Quit stalling.	لا تماطل.	توقف عن المماطلة.

4) استخدام اللُّغة العامية في مقابل اللُّغة الفصحى:

Sentence	Human translation	Google translation
You've cheated.	أنت مخادع	لقد غشيت.

5) الرُّتبة الخامسة ($0,6 \leq S.S < 0,7$)

جدول (11): عيِّنة من مخرجات (BERT) تنتمي للرتبة الخامسة وَفْق نسبة التَّشابه

M	Sentence	Human translation	Google translation	BERT
1	Close, oh Sesame	أقفل يا سمسم	قريب يا سمسم	0.6999
2	She kind of looks like Dooly.	تشبه دولي.	إنها تبدو مثل Dooly.	0.6995
3	I'm begging you	أتوسل إليك	أتوسل إليك	0.6953
4	Warehouse.	مكان للتخزين	مستودع.	0.6936
5	I will finish off Zero...!	سأقتل زيرو	سأنهي الصفر ...!	0.6877
6	they'll tell you.	سيبخرونك.	سيقولون لك.	0.6795
7	Stiff necked as ever.	صلبة الرأس كعادتك	رقبة شديدة من أي وقت مضى.	0.6728
8	I can't.	لا أستطيع	لا أستطيع.	0.6658
9	Because my head's in the clouds	هذا لأن شعري مبعثر	لأن رأسي في الغيوم	0.6558
10	Before the tides of passion Cool within you	قبل أن تخمد ثورتكن	قبل مد العاطفة بارد بداخلك	0.6554

الملاحظات والاستنتاجات

بدراسة جمل هذه المجموعة نجد أنَّه يمكن تقسيمها إلى فئات عدة:

1) أزواج متشابهة، ولذا كان من المفترض أن ترد في جمل الرتبة الأولى، مثل:

Sentence	Human translation	Google translation
I'm begging you	أتوسل إليك	أتوسل إليك

2) أزواج متشابهة في الدِّلالة مختلفة في المبنى:

Sentence	Human translation	Google translation
Warehouse.	مكان للتخزين	مستودع.
You Look Familiar.	أنت تبدين مشهورة	أنت تبدو مألوفة.

3) أزواج متشابهة في الدِّلالة لكن (BERT)، صنَّفها ذات نسبة تشابه دلالي قليل بسبب الأخطاء اللُّغويَّة، مثل:

Sentence	Human translation	Google translation
they'll tell you.	سيبخرونك.	سيقولون لك.
I can't.	لا أستطيع	لا أستطيع.
Of course it's dead	بالطبع ميت	بالطبع مات

6) الرُّتبة السَّادسة ($0,5 \leq S.S < 0,6$)

جدول (12): عيّنة من مخرجات (BERT) تنتمي للرُّتبة السَّادسة وَفْق نسبة التَّشابه

M	Sentence	Human translation	Google translation	BERT
1	You'll be acquitted.	سيبرؤنك	ستتم تبرئتك.	0.5999
2	Queens	ملكات.	كوينز	0.5876
3	Get in there. No.	اجلس هنالا	ادخل هناك. رقم.	0.5733
4	Scared?	خائف؟	مفزوع؟	0.5641
5	Yes. I'm a composer.	أنني مؤلف	نعم. أنا ملحن.	0.5587
6	Andrea	.	أندريا	0.5503
7	Yes.	كيف ممتع جدا.	نعم.	0.5380
8	Turn back	ارجعوا إلى الشاطئ	تراجع	0.5350
9	Throw him off	نقذفة.	ارميه بعيدا	0.5222
10	Are you working overitime?	هل تعمل ساعات إضافية؟	هل تعمل في البحر؟	0.5927

الملاحظات والاستنتاجات

بالنَّظر في جمل هذه المجموعة يمكن تصنيف أزواجها في العناصر التَّالية:

- جمل ذات دلالات مختلفة.
- جمل متقاربة في الدِّلالة ولكن نسبة التقارب قليلة.
- اختلاف الدِّلالة بين زوجي الجملة بسبب الأخطاء الإملائيَّة.
- اختلاف الكتابة الإملائيَّة بين زوجي الجملة.
- وجود كلمات في الجملة العربيَّة ليست موجودة في الجملة الإنجليزيَّة، فأدى إلى تباعد نسبة التشابه الدلالي بين الجملتين.

7) الرُّتبة السَّابعة ($0,4 \leq S.S < 0,5$)

جدول (13): عيِّنة من مخرجات (BERT) تنتمي للرُّتبة السَّابعة وَفْق نسبة التَّشابه

M	Sentence	Human translation	Google translation	BERT
1	You drive.	قم انت بالقيادة	أنت تسوق.	0.4977
2	What are you squawking about?	لماذا تزعقى؟	عن ماذا تصرخ؟	0.4610
3	Everybody's doing it.	الجميع يفعل ذلك	الجميع يفعل ذلك.	0.4987
4	Nocturnal conferences are bad for the nerves.	المؤتمراتالليلية ليست في صالح الأعصاب	المؤتمرات الليلية ضارة بالأعصاب.	0.4987
5	Old people.	العجائز	كبار السن.	0.4167
6	Isamu	جيد ولكن ما الذي تدخر لأجله؟	Isamu	0.4233
7	The artist?	الممثلة؟	الفنان؟	0.4232
8	Da da da ba bum barump	عديم الفائدة	Da da da ba bum barump	0.4123
9	HelpSerpentSerpent	طرقك أيتها الملكة	ساعدني ثعبان	0.4063
10	You got all dirty again.	لقد أتسخت مجددا	لقد أصبحت كل قذرة مرة أخرى.	0.4001

الملاحظات والاستنتاجات

أعطى (BERT) أزواج جمل هذه المجموعة نسبة تشابه أقل من 50%؛ نظرًا لكثرة المشكلات فيها، وتعود تلك المشكلات إلى استخدام العامية مقابل الفصحى، كما في:

Sentence	Human translation	Google translation
You drive.	قم انت بالقيادة	أنت تسوق.
What are you squawking about?	لماذا تزعق؟	عن ماذا تصرخ؟

ومن المآخذ التي ظهرت على (BERT) في هذه المجموعة أنَّه أيضًا صنَّف جملًا متطابقة في الدِّلالة والمبنى ضمن نسب التَّشابه القليلة، وهذا تصنيف خاطئ، كما في:

Sentence	Human translation	Google translation
Everybody's doing it.	الجميع يفعل ذلك	الجميع يفعل ذلك.

أيضًا وجود جمل تعطي الدِّلالة نفسَها لكن تركيبها مختلف، وهو ما يستطع (BERT) التنبه له، كما في:

Sentence	Human translation	Google translation
Nocturnal conferences are bad for the nerves.	المؤتمراتالليلية ليست في صالح الأعصاب	المؤتمرات الليلية ضارة بالأعصاب.
Old people.	العجائز	كبار السن.
The artist?	الممثلة؟	الفنان؟
He's a psychiatrist.	انه نفسانى…	إنه طبيب نفساني.

إلى جانب ذلك أيضًا وجد أزواج جمل متقاربة إلى حد ما، الفارق بينها في التَّعريف والتَّنكير كما في:

Sentence	Human translation	Google translation
Brute force.	قـوة غـاشمة	القوة الغاشمة.

8) الرُّتبة الثَّامنة ($0,3 \leq S.S < 0,4$)

جدول (14): عيِّنة من مخرجات (BERT) تنتمي للرتبة الثَّامنة وَفْق نسبة التَّشابه

M	Sentence	Human translation	Google translation	BERT
1	Something has happened.	لقد حدث شئ ما	لم يحدث شيء.	0.3888
2	Bingo	2.. 3	بنغو	0.3888
3	I want you to meet a very dear...	....	أريدك أن تقابل عزيزا جدا...	0.3876
4	Promised me my parole.	أعطوني كلمة شرف	وعدني بالإفراج المشروط.	0.3805
5	He's bats.	إنه مجنون	إنه خفافيش.	0.3804
6	Steering wheel?	عجلة القيادة؟	المقود؟	0.3754
7	I insist.	أنا مصممة	أنا أصر.	0.3459
8	I haven't got a mother.	ليسلديام.	ليس لدي أم.	0.3153
9	How about looking for the exit?	لم لا تبحثين عن باب الخروج؟	ماذا عن البحث عن المخرج؟	0.3481
10	I wanna pack it.	أنا اريد تغليفها.	أريد أن أحزمه.	0.3099
11	Go home.	احصل على افادات من كلا منهما	اذهب للمنزل	0.3054
12	Wait a minute.	نحن لسنا مشغولون	انتظر دقيقة.	0.3010

الملاحظات والاستنتاجات

1) أكثر أزواج الجمل في هذه الرتبة لا علاقة بينها إطلاقًا سواء في الدِّلالة أو المعنى، وبالتَّالي فإنَّ نسبة التَّشابه بينها من المفترض أن تكون (صفرًا).

2) تضمنت هذه الرُّتبة عددًا قليلًا من الجمل التي تحمل معنى مجازيًّا، أو اصطلاحيًّا، مثل التَّعبير (He's bats.)؛ حيث إنه في التَّرجمة البشريَّة بمعنى

(إنه مجنون)، أما في التَّرجمة الآليَّة فإنَّ التَّرجمة الآليَّة ترجمته ترجمة حرفية بمعنى (إنه خفافيش)، وبالبحث عن معاني كلمة (Bats) ظهر أنها تستخدم بمعنى (مجنون)[7]، ولكن جوجل لا يدرك هذا الاستخدام، فترجمه ترجمة حرفيَّة.

3) تضمنت عددًا قليلًا من الجمل المختلفة في المبنى القريبة نسبيًّا في الدِّلالة، مثل (.I insist)؛ فكانت التَّرجمة البشريَّة لها (أنا مصممة)، فيما كانت التَّرجمة الآليَّة لها (أنا أصر)، وكلاهما قريب في المعنى، واختلاف الدِّلالة بسيط بين التذكير والتَّأنيث، وأيضًا ينم عن عدم إدراك (BERT) لظاهرة المترادفات؛ حيث إن (أصر) و(مصمم) لهما الدِّلالة نفسها.

4) هناك أيضًا مجموعة من الجمل التي تتطابق في المعنى والمبنى، ولكن نظرًا للأخطاء الكتابيَّة وأخطاء النسخ وردت بصورة متشابكة في التَّرجمة اليدويَّة، وقد وضعها (BERT) في هذه الرُّتبة رغم أن من الصحيح وضعها في الرُّتبة الأولى؛ نظرًا لنسبة التَّشابه الكبيرة بين زوجي الجملة.

ويمكن اقتراح آليَّة لمساعدة (BERT) على اكتشاف العلاقة بين أزواج الجمل في حالة تشابك كلمات إحدى الجملتين من خلال تدريبه على عدد من الأمثلة وتزويده بالسِّمات التي قد يكتشف من خلالها التقارب بين الجملتين.

5) من الجمل التي بينها تشابه دلالي أيضًا:

Sentence	Human translation	Google translation	BERT
How about looking for the exit?	لم لا تبحثين عن باب الخروج؟	ماذا عن البحث عن المخرج؟	0.3481

فكلا الترجمتين لهما الدلالة نفسها مع اختلاف في أسلوب الخطاب في التَّرجمة اليدوية نظرًا لأنها معتمدة على السياق والحوار الدائر بين شخصين.

6) اختيار الآلة للتَّرجمة المصطلحيَّة دون مراعاة المقابل الدارج، كما في:

Sentence	Human translation	Google translation
Steering wheel?	عجلة القيادة؟	المقود؟

فكلا التَّعبيرين صحيح، لكن هناك استخدام دارج ربما تجهله الآلة، وهو (عجلة القيادة)؛ لذا فقد صنفها (BERT) على أن زوجي الجملة غير متطابقان.

7) كما أنَّ هذه المجموعة تضمنت عددًا من الجمل التي تعطي يكون معنى إحدى الجملتين عكس دلالة الجملة الأخرى، كما في:

Sentence	Human translation	Google translation
Something has happened.	لقد حدث شئ ما	لم يحدث شيء.

فالملاحظ على جمل هذه الرتبة، أن النسبة الأكبر من جملها لا علاقة بينها سواء على مستوى الدلالة أو المبنى والتركيب، فيما عدا نسبة قليلة من الجمل المجازية، أو الجمل التي لم يستطع (BERT) تمييز تطابقها بسبب أخطاء النسخ والكتابة، وكذلك الرتب الأقل منها، لأنها بطبيعة الحال لن تكون أفضل منها في التشابه أو الموازاة بين أزواج الجمل.

نتائج الدِّراسة

أ. ساعد هذا المسح على تقييم مخرجات (BERT) عند تعامله مع اللُّغة العربيَّة، اعتمادًا على نسبة الجمل التي استطاع تصنيفها وفق درجة تشابه زوجيها؛ فرغم أن كفاءة الأداة مع اللغة الإنجليزية بلغت حوالي (65%)، وهو ما يرجع إلى تمرُّس تعامل الأداة مع اللُّغة الإنجليزيَّة والتدرُّب على عدد لا حصر له من النُّصوص المختلفة، في مقابل (40%) مع اللُّغة العربيَّة؛ نظرًا لطبيعتها، بالإضافة إلى أنَّها تأتي ضمن حزمة من اللُّغات التي يعالجها (BERT)؛ والتي تخطى عددها 50 لغة، بينها الصِّينيَّة والفرنسيَّة والإسبانيَّة والرُّوسيَّة، وغيرها[8].

ب. تمَّ الوقوف على المشكلات التي تعيق عمل (BERT) وأدَّت إلى ضعف دقة مخرجاته عند التَّعامل مع اللُّغة العربيَّة، وفيما يلي نستعرض هذه المشكلات ممثلين لها ببعض الأمثلة السريعة.

1) عدم قدرته على اكتشاف أزواج الجمل التي تعطي الدِّلالة نفسها؛ نظرًا لاعتماده بصورة كبيرة على شكل الجملة وتركيبها ومفرداتها:

جدول (15): (BERT) غير قادر على اكتشاف الجمل التي تعطي الدلالة نفسها بسبب شكلها

M	Sentence	Human translation	Google translation	BERT
1	All we need is the briefcase.	كل ما نحتاج اليه هى الحقيبه	كل ما نحتاجه هو الحقيبة.	0.4913
2	I made no remark.	لا تعليق	لم أبدي أي ملاحظة.	0.4903
3	This is no time for secrets.	لم يعد هناك وقت للأسرار	هذا ليس وقت الأسرار.	0.4717
4	Steering wheel?	عجلة القيادة؟	المقود؟	0.3754
5	Postponed.	أجلها؟	مؤجل.	0.3627
6	What isperjury ?	ما هي شهادة الزور؟	ما هو الحنث باليمين؟	0.3605

2) لا يتعامل بصورة موحَّدة مع عدد من أنماط الأخطاء، مثل الجمل التي تواجه رمزًا معينًا، كان من الأولى أن يصنِّفها كلَّها تحت نسبة تشابه معيّنة؛ لأنَّ الرَّمز لم يختلف، أو أنَّه جملة في مقابل رمز، بل إنه أعطاها نسبة تشابه أكبر من أزواج جمل بينها تشابه فعلًا، كما الجدول (13).

جدول (16): (BERT) لا يتعامل مع الخطأ نفسه بصورة موحدة

M	Sentence	Human translation	Google translation	BERT
1	Sometimes...	....	بعض الأحيان...	0.6465
2	Andrea	.	أندريا	0.5503
3	I don't	.	أنا لا	0.5374
4	Then we've got something in common.	.	لذلك لدينا شيء واحد مشترك	0.5348
5	But, Doctor, I...	....	لكن، دكتور، أنا...	0.5265
6	I'm sick, Joe.	..	أنا مريض يا جو.	0.3173

3) مشكلة التَّعريف والتَّنكير التي تؤثِّر بشكل كبير على نسبة التشابه بين أزواج الجمل، حتى وإن وُجد اختلاف كان من الأفضل أن تكون نسبة الاختلاف بسيطة وليست بهذا الفارق، كما في:

جدول (17): مشكلة التَّعريف والتَّنكير تؤثِّر على مخرجات (BERT)

M	Sentence	Human translation	Google translation	BERT
1	Brute force.	قـوة غـاشمة	القوة الغاشمة.	0.4482
2	Tension?	التوتر؟	توتر؟	0.4127

4) إعطاؤه نسبة تشابه بين جملتين لهما ليس دلالتهما مختلفة، بل إن كل جملة لها دلالة عكس دلالة الجملة الأخرى، كما في:

جدول (18): (BERT) أعطى نسبة تشابه بين جملتين متضادتين

Sentence	Human translation	Google translation
Something has happened.	لقد حدث شيء ما	لم يحدث شيء.

5) كيف يمكن إعطاء نسبة تشابه بين جملتين لا علاقة بينهما على الإطلاق؟ بل إنَّه لا يعطي لها نسبة تشابه قليلة، على العكس إنَّ نسبة التَّشابه تكون عاليَّة، ولا وجود لأي وجه شبه بين الجملتين، كما في:

جدول (19): (BERT) يعطي نسبة تشابه بين جمل لا علاقة بينها إطلاقًا

M	Sentence	Human translation	Google translation	BERT
1	he's always got his head in the clouds	هو دائما شارد الذهن هكذا	دائما ما يكون رأسه في الغيوم	0.7876
2	Take the clouds from your eyes.	أبعد الغشاوة عن عينيك.	خذ الغيوم من عينيك.	0.7735
3	Close, oh Sesame	أقفل يا سمسم	قريب يا سمسم	0.6999
4	Honey.	عزيزتي	عسل.	0.6944
5	Follow that coach	أتبعوا العربة	اتبع هذا المدرب	0.5997

6) تأثير الأخطاء الإملائيَّة على مخرجاته، بل إنَّه يعطي للخطأ الإملائي نفسه نسب تشابه مختلفة، كما في:

جدول (20): تأثير الأخطاء الإملائيَّة على مخرجات (BERT)

M	Sentence	Human translation	Google translation	BERT
1	You're a cop	أنتشرطي	أنت شرطي	0.4191
2	It's the dog	إنهالكلب	إنه الكلب	0.5982
3	Britannia	بيرطانيا	بريتانيا	0.6450
4	Who would dare rob the betrothed of the great Khan?	من الذي يجرؤ على سرقة خطيبة خان المعظم؟	من يجرؤ على سرقة خطيب الخان العظيم؟	0.7893

7) وجود بعض أزواج الجمل المتطابقة، لكنه يعطيها نسب تشابه قليلة مقارنة بدرجة تشابهها، كما في:

جدول (21): (BERT) يعطي أزواج جمل متطابقة نسبة تشابه قليلة

M	Sentence	Human translation	Google translation	BERT
1	I'm begging you	أتوسل إليك	أتوسل إليك	0.6953
2	She's reading a magazine.	إنها تقرأ مجلة	إنها تقرأ مجلة.	0.8281
3	It's the dog	إنهالكلب	إنه الكلب	0.5982
4	Open up	افتحي	افتح	0.6295
5	And in your heart?	1972وفي قلبك؟	وفي قلبك؟	0.6378

8) وجود جمل تكاد تكون متطابقة، لكنه يعطيها نسب تشابه قليلة، كما في:

جدول (22): وجود أزواج جمل متطابقة بنسبة كبيرة و(BERT) يعطيها نسبة تشابه قليلة

M	Sentence	Human translation	Google translation	BERT
1	Stay in bed.	لازم الفراش	ابق في الفراش.	0.6051
2	Scared?	خائف؟	مفزوع؟	0.5641
3	He's a spy.	إنه واشي	إنه جاسوس.	0.5379
4	Where's the necklace?	اين العقد؟	أين القلادة؟	0.5171
5	I did not care.	انا لا اعبأ	لم أهتم.	0.5075

9) كما أمكن للبحث توفير نموذج لُغويٍّ لتدريب (BERT) على التَّعامل مع اللُّغة العربيَّة بصورة أكثر دقة، وهو ما يتضمنه ملحق البحث.

التوصيات

بناء على هذا العرض يمكننا اقتراح بناء مدوَّنة متوازية تتضمن أنماط الجمل التي أخطأ (BERT) في تقييم درجة التشابه بينها، ثم تحكيمها يدويًّا من قبل عدد من خبراء اللُّغة؛ لأعطائها درجة التَّشابه الحقيقيَّة، ثم تدريب الآلة عليها، لتتمكن من التَّعامل معها، وبالتالي التَّعامل مع الأنماط المشابهة لها، والتغلب على مثل هذه المشكلات.

المراجع

(1) (لبيب) محمد مجدي: "موارد الترجمة الآلية بين اللغة العربية والإنجليزية.. معالجة لغوية حاسوبية"، رسالة دكتوراه، كلية الآداب، جامعة عين شمس، ص142

(2) https://translate.google.com

(3) لمزيد من المعلومات انظر:

https://www.wikiwand.com/en/BERT_(language_model)

And See: https://www.techopedia.com/definition/34116/bidirectional-encoder-representations-from-transformers-bert

(4) See, "Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching", Op.cit, p.324.

(5) **Devlin, J. et al**: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805v2 [cs.CL] 24 May 2019.

(6) https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1

(7) https://www.wordreference.com/enar/Bats

(8) https://www.sbert.net/docs/pretrained_models.html