الدراسة المرجعية حولParaphrase Identification

مقدمة عامة :

- تُعد Paraphrase Identification أو التعرف على الجمل المتشابهة واحدة من أهم المهام في مجال معالجة اللغات الطبيعية (NLP) ، حيث تهدف إلى تحديد ما إذا كانت جملتان مختلفتان في الصياغة تحملان نفس المعنى. تُستخدم هذه المهمة في العديد من التطبيقات العملية مثل:
- **كشف السرقة الأدبية** (Plagiarism Detection)
- **الترجمة الآلية** (Machine Translation)
- **تلخيص النصوص** (Text Summarization)

يعود التحدي الأساسي في **Paraphrase Identification** إلى تنوع أساليب التعبير البشري، إذ يمكن صياغة نفس الفكرة باستخدام كلمات أو تراكيب نحوية مختلفة تمامًا. على سبيل المثال:

- **الجملة الأولى**: "The cat is sleeping on the couch."
- **الجملة الثانية**: "The feline is napping on the sofa."

على الرغم من اختلاف الكلمات مثل "cat" و "feline"أو "sleeping" و"napping"، إلا أن المعنى في الجملتين متشابه.

وأيضا لا تزال هناك عدة تحديات تواجه الباحثين في هذا المجال، منها:

1. **فهم السياق الضمني** :بعض الجمل تتطلب معرفة الخلفية الثقافية أو السياقية.
2. **التعبيرات الاصطلاحية** :التعامل مع العبارات التي تحمل معاني غير حرفية.
3. **تنوع اللغة** :وجود مرادفات متعددة وصيغ تركيبية مختلفة تؤدي إلى نفس المعنى.

التطور التقني في المجال

مع التقدم في تقنيات التعلم الآلي التقليدي والتعلم العميق، تم تطوير العديد من الأساليب التي تُعالج هذه المهمة. حيث اعتمدت الأساليب التقليدية على التمثيلات الإحصائية للنصوص مثل TF-IDFو N-grams، بينما جاءت النماذج الحديثة مثل BERTو Transformer-based Modelsلتقدم أداءً متفوقًا بفضل قدرتها على فهم السياق الكامل للنص.

- **المنهجيات التقليدية(Traditional Methods):**

- الأساليب الإحصائية(Statistical Methods) :

تعد الأساليب الإحصائية من أقدم الطرق المستخدمة في **Paraphrase Identification**، وهي تعتمد على حساب درجة التشابه بين الجمل باستخدام تمثيلات النصوص الإحصائية. من أبرز هذه الأساليب:

1. **TF-IDF (Term Frequency-Inverse Document Frequency) :**

يُعتبر TF-IDF أحد الأساليب البسيطة والأكثر استخدامًا في تحليل النصوص وتمثيل الكلمات. يُستخدم TF-IDF لحساب أهمية الكلمات داخل جملة أو مستند معين بناءً على تكرار ظهور الكلمة (TF) في المستند الواحد ونسبة وجودها في مستندات أخرى (IDF) يتم استخدام هذه الميزة لتحديد مدى أهمية الكلمة بالنسبة للجملة.

آلية عمل :TF-IDF

1. **تواتر الكلمة (TF) :** يتم حساب عدد مرات ظهور الكلمة في الجملة أو المستند، ويتم قياسه باستخدام الصيغة :

$$\mathrm{TF}(t) = \frac{\text{عدد مرات ظهور الكلمة } t \text{ في المستند}}{\text{عدد الكلمات في المستند}}$$

2. **تواتر المستند المعكوس (IDF) :** يتم حساب مقدار أهمية الكلمة عبر المستندات تُحسب باستخدام الصيغة :

$$\mathrm{IDF}(t) = \log\left(\frac{N}{df(t)}\right)$$

حيث:

- NNN هو إجمالي عدد المستندات في المجموعة.
- df(t) هو عدد المستندات التي تحتوي على الكلمة t.

3. **حساب الـ TF-IDF :**

$$\text{TF-IDF}(t) = \mathrm{TF}(t) \times \mathrm{IDF}(t)$$

المزايا:

- **بسيط وسهل التنفيذ** : يمكن حسابه بسرعة وبدون الحاجة إلى موارد حسابية ضخمة.
- **لا يعتمد على السياق** : مناسب لبيانات صغيرة أو لغات مُهيكلة.

العيوب:

- **عدم القدرة على التقاط العلاقات السياقية** : في حالات معينة قد تحتوي الجمل على كلمات ذات معاني مختلفة بناءً على السياق، ولا يستطيع **TF-IDF** فهم هذه العلاقات.
- **التأثير الكبير للكلمات الشائعة** : الكلمات الشائعة في اللغة مثل "of" ,"is" ,"the" تحصل على **أوزان عالية** لأنها تظهر بكثرة في العديد من المستندات، مما يقلل من دقة تمثيل الكلمات ذات المعنى الفعلي.

مثال:

1: "The cat sat on the mat"
2: "The feline sat on the rug"

- **كلمة "cat"** و **"feline"** تعتبر كلمات مترادفة ولكن **TF-IDF** قد يعطيهما أوزانًا مماثلة على الرغم من اختلاف السياق في الجملة.

2. N-grams (N-gram Models) :

تعتمد نماذج N-grams على تقسيم النص إلى تسلسل من N كلمات متتابعة. يتم بعد ذلك مقارنة الأنماط المتكررة أو التسلسلات بين الجمل باستخدام مقاييس تشابه مثل Jaccard Similarity أو Cosine Similarity يشير N-gram إلى عدد الكلمات المتتالية التي يتم أخذها في الاعتبار في النموذج.

- **1-gram (unigram)** كل كلمة مفردة.
- **2-gram (bigram)** تسلسل مكون من كلمتين.
- **3-gram (trigram)** تسلسل مكون من ثلاث كلمات.

آلية العمل:

1. تقسيم الجملة إلى N-grams: على سبيل المثال، إذا كانت الجملة "The cat sat on the mat" هي الجملة التي نعمل عليها:
    - **Unigrams (1-gram)**: ["The", "cat", "sat", "on", "the", "mat"]
    - **Bigrams (2-grams)**: ["The cat", "cat sat", "sat on", "on the", "the mat"]
    - **Trigrams (3-grams)**: ["The cat sat", "cat sat on", "sat on the", "on the mat"]
2. مقارنة التكرارات: تُقارن الجمل بناءً على عدد N-gramsالمتكررة بين الجمل المختلفة.

المزايا:

- **مناسب للبيانات التي تحتوي على تسلسل**: مثل الجمل التي تحتوي على ترابط منطقي بين الكلمات المتتالية.
- **مرونة في معالجة التركيبات النحوية**: يمكن تحليل النمط التركيبي للكلمات عبر الأنماط المتتالية.

العيوب:

- **الحاجة إلى حجم كبير من البيانات** :النموذج يحتاج إلى عدد كبير من **الأنماط (N-grams)** للتمثيل بدقة.
- **الصعوبة في التعامل مع مرادفات** :عندما تتغير الكلمات ولكن المعنى يبقى كما هو، **N-grams** لا يمكنه التعامل مع هذه التغييرات.

مثال:

1"The cat sat on the mat" :
2"The feline sat on the rug" :

- **نموذج الـ Bigrams قد يعطي التشابه بين "sat on" في كلا الجملتين، لكنه قد يفشل في التقاط اختلاف "cat" و "feline" و "mat" و "rug" بسبب اختلاف المفردات رغم أن المعنى مشابه.**

- الأساليب المعتمدة على القواعد(Rule-based Methods) :

المنهجيات المعتمدة على القواعد تُعتبر من أقدم الطرق المستخدمة في معالجة اللغات الطبيعية، حيث تعتمد على تحديد التركيبات اللغوية أو الأنماط التي تشير إلى وجود تشابه بين الجمل بناءً على قواعد لغوية محددة مسبقًا. في هذا السياق، يمكن تحديد الكلمات أو العبارات التي تحمل معاني مشابهة باستخدام **القواميس اللغوية** و **المراجع** اللغوية، مثل **القواميس المعجمية** (Lexical Databases)و **قواميس المرادفات** (Thesaurus).

- **التمثيلات النحوية (Syntactic Representations) :**
  - تعتمد هذه الطريقة على **تحليل التراكيب النحوية** (Parsing)للجمل باستخدام أدوات مثل **شجرة التراكيب (Parsing Trees)**الهدف من تحليل التراكيب النحوية هو فهم **العلاقات بين الكلمات** في الجملة وكيفية تركيبها في سياق معنوي. بمجرد بناء الشجرة النحوية للجملتين، يتم مقارنة **التراكيب النحوية** بين الجمل لتحديد إذا كان التشابه قائمًا أم لا.

الخطوات الرئيسية في هذه الطريقة:

1. **تحليل الشجرة النحوية:(Parsing Trees)**
   - يتم تحويل الجملة إلى شجرة نحوية باستخدام **محللات النحو (Parsers)** على سبيل المثال، يمكن استخدام **محللات النحو القاعدية (Context-Free Grammar)** لتحليل الجمل بناءً على قواعد نحوية محددة مسبقًا.
   - بناءً على هذه التحليلات، يتم استخراج الهيكل النحوي الذي يوضح كيفية ترابط الكلمات داخل الجملة.
2. **مقارنة التراكيب النحوية:**
   - بعد الحصول على الشجرة النحوية لكل جملة، يمكن مقارنة **التركيبات النحوية** بين الجمل باستخدام **مسافة Levenshtein** أو **مقاييس التشابه بين الأشجار** مثل **Tree Edit Distance**

- إذا كانت الأشجار النحوية للجملتين متشابهة بدرجة كبيرة، فهذا يشير إلى أنهما تحملان معاني مشابهة.

3. **استخدام القواميس والمراجع اللغوية:**

- في هذه المنهجية، تُستخدم القواميس مثل **WordNet** أو **FrameNet** لربط الكلمات في الجمل بمعانيها الدقيقة وتحديد المرادفات، مما يعزز من دقة تحديد التشابه بين الجمل.

العيوب:

1. **الحاجة إلى قواميس لغوية دقيقة:**

- يعتمد هذا الأسلوب بشكل كبير على القواميس اللغوية، مثل **WordNet**أو **FrameNet**، التي توفر معلومات حول المعاني المتعددة للكلمات. في حال كانت هذه القواميس ناقصة أو غير دقيقة، فإن التشابه بين الجمل قد لا يتم تحديده بشكل صحيح.

- القواميس قد تفتقر إلى تغطية المعاني الضمنية أو العبارات الاصطلاحية التي لا تعكس بشكل دقيق المعاني الحقيقية للجمل.

2. **التعقيد الحسابي:**

- تحليل التراكيب النحوية يتطلب عمليات حسابية معقدة. الترجمة إلى أشجار نحوية وتحليلها يحتاج إلى موارد حوسبية كبيرة، خصوصًا في حالة الجمل الطويلة أو المتشعبة.

- بالإضافة إلى ذلك، قد تتطلب هذه العمليات وقتًا أطول عند التعامل مع بيانات ضخمة.

3. **التعامل المحدود مع السياق الكامل:**

- على الرغم من أن هذه الطريقة تحلل الجمل بشكل نحوي، إلا أن فهم السياق الكامل للجمل يبقى محدودًا. على سبيل المثال، قد لا تتمكن من فهم الجمل التي تحتوي على التلميحات الثقافية أو المفاهيم غير المباشرة.

- يمكن لهذه الطريقة أن تُخطئ في تحديد التشابه في الحالات التي تحتوي فيها الجمل على تعبيرات غير حرفية أو تلاعب لغوي مثل السخرية.

4. **نقص التكيف مع التغييرات في الأسلوب أو اللغة:**

- في حالات الاختلافات اللغوية أو الأسلوبية، قد لا تكون القواعد اللغوية المحددة مسبقًا قادرة على التكيف مع هذه التغييرات. على سبيل المثال، إذا كان هناك تغيير طفيف في تركيب الجملة أو إذا تم استخدام مرادف غير مألوف، قد يفشل النظام في تحديد التشابه.

- **استخدام القواعد اللغوية لتمثيل العلاقات المعنوية(Lexical and Semantic Rules)**

في بعض الأساليب المعتمدة على القواعد، يتم تحديد المرادفات و العلاقات الدلالية بين الكلمات بناءً على القواميس و المراجع اللغوية .تعتمد هذه الطريقة على مقارنة الكلمات في الجمل باستخدام المرادفات مثل "synonyms" "hypernyms" و العلاقات بين الكلمات مثل Antonyms.

أدوات مثل **WordNet** توفر هذه العلاقات الدلالية، مما يسمح للموديل بتحديد التشابه بين الكلمات حتى لو كانت الكلمات نفسها تختلف، على سبيل المثال، يمكن أن يعرف النموذج أن كلمة "happy" مرادفة لـ "joyful" .

- **المزايا** :تعمل هذه الطريقة بشكل جيد مع الكلمات ذات المعاني المباشرة (مثل الكلمات ذات المرادفات الواضحة).

- **العيوب** :قد تكون العلاقات المعنوية التي يتم استخراجها ضحلة وغير دقيقة في الحالات التي تحتوي على معاني غير مباشرة أو جديدة

• المنهجيات المعتمدة على التعلم الآلي التقليدي (Traditional Machine Learning Methods)

- مع تطور تقنيات التعلم الآلي التقليدي، تم استخدام نماذج تعتمد على الخصائص اليدوية Hand-crafted Features المُستخرجة من النصوص، مثل **TF-IDF**، **Bag of Words (BoW)**، و**N-grams** تعتمد هذه النماذج على تمثيل النصوص كمتجهات عددية ضمن فضاء رياضي يمكن للنماذج التعامل معه لتصنيف النصوص المتشابهة.

: Naive Bayes Classifier

- **Naive Bayes** هو نموذج تعلم آلي إحصائي يعتمد على قانون بايز لتصنيف النصوص إلى فئات.
- يُفترض في هذا النموذج أن جميع الكلمات في النص مستقلة عن بعضها البعض (فرضية الاستقلال المشروط)، وهو ما يجعله "naive" .

آلية العمل:

1. يتم حساب الاحتمالات الشرطية لكل كلمة في النص بناءً على **تكرارها** في البيانات.
2. يُستخدم قانون بايز لحساب احتمالية أن تنتمي الجملة إلى فئة معينة (مشابه أو غير مشابه):

$$\frac{P(X|C) \cdot P(C)}{P(X)} = P(C|X)$$

حيث:

- P(C|X)الاحتمال المُرجح لانتماء النص إلى الفئة C بناءً على البيانات X.
- P(X|C)الاحتمال الشرطي للكلمات داخل الفئة.
- P(C)الاحتمال المسبق للفئة.
- P(X)الاحتمال الإجمالي للكلمات.

3. يتم اختيار الفئة ذات أعلى احتمال بناءً على النص المُدخل.

المزايا:

- بسيط وسريع، يُعد من أسهل النماذج في التدريب والتنفيذ.
- فعّال مع البيانات النصية الصغيرة والمتوسطة الحجم.

العيوب:

- **فرضية الاستقلال:** يعتمد على فرضية أن الكلمات مستقلة عن بعضها البعض، وهو ما لا ينطبق في النصوص الطبيعية.
- **إهمال السياق:** لا يستطيع التعامل مع العلاقات السياقية بين الكلمات.

• المنهجيات المعتمدة على التعلم العميق (Deep Learning Methods)

الشبكات العصبية التكرارية (RNNs) و LSTMs :

- تعتمد **الشبكات العصبية التكرارية (Recurrent Neural Networks)** على معالجة البيانات المتسلسلة مثل النصوص والجمل كلمةً بكلمة، مع الحفاظ على المعلومات المُخزّنة من الكلمات السابقة عند تحليل الكلمات اللاحقة.
- بينما نموذج **LSTM (Long Short-Term Memory)** كتحسين على **RNNs** لحل مشكلة النسيان (Vanishing Gradient Problem)في الجمل الطويلة، حيث يحتفظ بـ الذاكرة الطويلة والقصيرة المدى.

آلية العمل:

1. **إدخال النص على شكل تسلسل:** يتم تمرير الجملة كلمة بكلمة إلى الشبكة.
2. **الحفاظ على المعلومات عبر الزمن:** تحتفظ وحدات **LSTM** بمعلومات عن الكلمات السابقة باستخدام بوابات (Gates) مثل:
    - **بوابة الإدخال (Input Gate)** تحدد أي جزء من المدخلات الجديدة سيتم تخزينه.
    - **بوابة النسيان (Forget Gate)** تحدد أي جزء من الذاكرة سيتم تجاهله.
    - **بوابة الإخراج (Output Gate)** تُحدد القيم التي ستُستخدم كخارج للنموذج.
3. **تمثيل الجملتين** :في مهمة **Paraphrase Identification**، يتم إدخال الجملتين إلى نموذج LSTM للحصول على تمثيلات نهائية لكل جملة.

المزايا:

- القدرة على معالجة الجمل الطويلة.

- فهم العلاقات المعقدة بين الكلمات بفضل البوابات في LSTM .

العيوب:

- بطء التدريب : يتطلب معالجة الكلمات بشكل تسلسلي، مما يزيد من زمن التدريب.
- صعوبة في فهم المعاني الضمنية : قد لا يتمكن النموذج من فهم الكلمات ذات العلاقات غير المباشرة.
- يتأثر بأداءه عند التعامل مع الجمل المتباعدة في السياق.

شبكات Transformer ونموذج BERT :

تعريف:

- يُعد **Transformer** من أبرز الابتكارات في التعلم العميق لمعالجة النصوص قدمه **Vaswani et al. (2017)** كبديل للشبكات التكرارية لحل مشكلات الكفاءة وعدم قدرة RNN على معالجة النصوص بالتوازي.
- يعتمد **Transformer** على آلية **Attention** التي تسمح للنموذج بالتركيز على الكلمات الأكثر أهمية لفهم السياق.
- نموذج **BERT (Bidirectional Encoder Representations from Transformers)** هو تطوير على Transformer ، حيث يقوم بتحليل النص بشكل ثنائي الاتجاه لفهم السياق الكامل.

آلية العمل في BERT :

1. **تمثيل الكلمات :** يتم تحويل الكلمات إلى تمثيلات مُدمجة (Embeddings) عبر طبقات Transformer.
2. **آلية Attention :** تعتمد على حساب الأهمية النسبية للكلمات في الجملة عبر معادلة:

$$Softmax\left(\frac{^{T}QK}{\sqrt{_{k}d}}\right)V = Attention(Q,K,V)$$

حيث:

- Q المتجهات المُدخلة للاستعلام (Query)
- K المتجهات الرئيسية (Key)
- V المتجهات القيمية (Value)
- Kd طول المتجه.

3. **تمثيل الجمل :** يتم دمج الجملتين وتحليل العلاقة بينهما باستخدام BERT عبر طبقة تصنيف نهائية

المزايا:

- **فهم السياق الثنائي الاتجاه :** يقرأ BERT النص في الاتجاهين (من اليسار إلى اليمين والعكس).
- دقة عالية في تحديد التشابه بين الجمل، نظرًا لقدرة Attention على التركيز على الكلمات المهمة.
- يمكن تطبيقه مباشرة على**Paraphrase Identification** بدون الحاجة إلى استخراج ميزات يدوية.

العيوب:

- **استهلاك الموارد :** يتطلب BERT موارد حوسبية ضخمة للتدريب.
- **زمن تدريب طويل:** يحتوي على عدد هائل من المعاملات(Parameters) .

Sentence-BERT (SBERT) :

تعريف:

- قدم **Reimers et al. (2019)** نموذج **Sentence-BERT** كتحسين على **BERT** لمعالجة المهام التي تتطلب مقارنة الجمل بكفاءة أكبر.
- يُعتبر SBERT شبكة (Siamese Network)، حيث يتم إدخال الجملتين إلى نموذجين متوازيين من BERT، ثم يتم مقارنة التمثيلات النهائية باستخدام **Cosine Similarity**.

آلية العمل:

1. **إدخال الجملتين إلى SBERT :** يتم إدخال الجمل إلى نسختين متطابقتين من BERT
2. **استخراج التمثيلات النهائية :** يقوم كل نموذج بإخراج تمثيل ثابت لكل جملة.
3. **قياس التشابه :**يتم استخدام مقاييس مثل **Cosine Similarity** لحساب درجة التشابه بين التمثيلين.

المزايا:

- **كفاءة في المقارنة :** تحويل الجمل إلى تمثيلات ثابتة، مما يُسهل مقارنة الجمل.
- **سرعة في التنفيذ :** مقارنة التمثيلات المُستخرجة أسرع من معالجة الجمل من الصفر.
- أداء مُحسّن في المهام التي تتطلب قياس التشابه النصي.

العيوب:

- **استهلاك الذاكرة :** نظرًا لاعتماده على BERT ، فإنه يحتاج إلى ذاكرة كبيرة عند التعامل مع مجموعات بيانات ضخمة.
- **حجم النموذج :** ما زال النموذج كبيرًا ويتطلب موارد جيدة للتدريب والتشغيل.

المقارنة بين المنهجيات العميقة:

النموذج	المزايا	العيوب
RNNs / LSTMs	فهم العلاقات بين الكلمات في الجمل المتسلسلة	بطء التدريب وصعوبة التعامل مع السياق الطويل
BERT / Transformers	دقة عالية وفهم السياق ثنائي الاتجاه	استهلاك موارد حوسبية وزمن تدريب طويل
Sentence-BERT	كفاءة في مقارنة الجمل وسرعة التنفيذ	استهلاك ذاكرة كبير وحجم النموذج الضخم