



معالجة اللّغة الطبيعية: فهم وتوليد اللّغة الطبيعية
بناء الملخصات أنموذجًا

Natural Language Processing: NLU & NLG
Building Summaries as a Model

عمر مزدك¹، مختار بن جلول²

مخبر الدراسات النحوية واللغوية بين التراث والحداثة في الجزائر

¹ جامعة ابن خلدون، تيارت (الجزائر)، omar.mezeddek@univ-tiaret.dz

² جامعة ابن خلدون، تيارت (الجزائر)، mokhtar.bendjelloul@univ-tiaret.dz

ملخص:

تأتي هذه الورقة البحثية لإلقاء الضوء على جزئية مهمة من جزئيات معالجة اللغات الطبيعية، والتي تعتبر من الدراسات البينية التي تجمع بين علوم الحاسوب وعلوم اللغة البشرية، عن طريق توظيف جملة من التقنيات والأدوات الآلية التي يتم التعامل معها بواسطة الخوارزميات التي تنبني على القواعد المنطقية والرياضية و مبادئ النظريات اللغوية من أجل تحليل اللّغة البشرية الطبيعية عن طريق النّمدجة والصّورنة، وسينصبُّ تركيزنا كما أسلفنا سابقا على كيفية فهم اللّغة وتوليدها لتكون مدخلا للمعالجة الآلية للغة الطبيعية. لقد أصبحت معالجة اللّغات الطبيعية ترتبط ارتباطاً وثيقاً بالذكاء الاصطناعي، الذي أضحي لا مناص منه، خاصة وأنّ كلّ التقنيّات كالتحليل التلقائي، وتلخيص النصوص، والترجمة الآلية، وغيرها تقوم على أساسه، كما سنعتمد في بحثنا هذا على التلخيص التلقائي للنصوص كأنموذج للتطبيق.

كلمات مفتاحية: معالجة اللغة الطبيعية؛ فهم اللغة الطبيعية؛ توليد اللّغة الطبيعية؛ الصّورنة؛ الترميز؛ التلخيص الاستخراجي؛ التلخيص الإجرائي.

Summary:

This research paper aims to shed light on an important aspect of Natural Language Processing, which is considered an interdisciplinary study that combines computer science and human language sciences. It employs a set of techniques and automated tools that are handled by algorithms based on logical and mathematical rules, as well as principles from linguistic theories, to analyze natural human language through modeling and representation. Our focus, as previously mentioned, will be on how to understand and generate language as an entry point for automated processing of natural language. Natural Language Processing has become closely linked with artificial intelligence, which has become indispensable, especially since all techniques such as automatic analysis, text summarization, machine translation, and others are based on it. In this research, we will rely on automatic text summarization as a model for application.

Keywords: Natural Language Processing; Natural Language Understanding; Natural Language Generation; Representation; Encoding; Extractive Summarization; Procedural Summarization

1.1. مقدمة:

يُعدّ مجال المعالجة الآلية للغة الطّبيعيّة (Automated natural language processing) من المجالات والتحدّيات الكبيرة في علوم الحاسوب والدّكاء الاصطناعي؛ إذ تسعى إلى تمكين الحاسوب والآلة عموماً من معالجة اللّغة البشريّة معالجةً تُمكن من التعرّف على اللّغة في شكلها المكتوب والمنطوق، وذلك من خلال تحليل اللّغة وفهمها وتولييفها وتولييدها. ليس من السهل تمكين الآلة القيام بمثل هكذا مهام، ففهم اللّغة - الذي يقوم على عدة مستويات متشابكة ومعقدة دلاليّة كانت أو سياقيّة أو حتّى تداوليّة - أمرٌ صعب ناهيك عن توليد اللّغة وإجراء الحوار مع الآلة؛ وإن كُنّا نرى مثل هذا مُتجسّداً نوعاً ما؛ غير أنّه لم يُحقّق نجاحاً وفاعليّة على المستوى الذي يُمكن القول إنّ الآلة تفهم وتنتج لغة مُطابقةً للّغة البشريّة.

وإنّ هذه الأبحاث التي نشهدها في مجال الدّكاء الاصطناعي والتطوّر الكبير في برامجه قام أساساً على محاكاة اللّغة البشريّة عن طريق تحليلها وصورتها، وبالتالي الوصول إلى تعلّم الآلة استناداً لعمليات التحليل والفهم وتخزين التمثيلات الدلاليّة المختلفة ثم توليد اللّغة. وسنقدّم في هذه الورقة تصوّراً نظريّاً وتطبيقياً حول معالجة اللّغة الطّبيعيّة بقسميها: فهم اللّغة وتوليد اللّغة، من خلال تقديم عملٍ تطبيقيٍّ حول تلخيص النصوص، مُنطلقين من التّساؤلات الآتية:

- 1- ماذا نعني بمعالجة اللّغة الطبيعيّة وقسميها فهم اللّغة وتوليد اللّغة حاسوبياً؟ وما العلاقة التي تربط بين اللّغة البشريّة ولغة الآلة؟
- 2- كيف تتمّ صورنة اللّغة الطبيعيّة إلى لغة رياضيّة يفهمها الحاسوب؟
- 3- كيف يتمّ الاعتماد على أدوات فهم وتوليد اللّغة الطبيعيّة في تلخيص النصوص؟

وللإجابة على هذا التساؤل سننطلق أولاً من شرح وتعريف معالجة اللّغة الطبيعيّة وفهم وتوليد اللّغة، كما سنتطرق في خضمّ الحديث عن أهميّة صياغة النظرية اللّغوية المناسبة- إلى مفهوم الصّورنة، وسنحاول تقديم عمل تطبيقي عن معالجة اللّغة الطبيعيّة في مجال تلخيص النصوص لتوضيح سير عمليّة المعالجة.

إنّ الهدف من هذه الورقة تقديم نظرة قريبة حول معالجة اللّغة الطبيعيّة، والمزج بين الجانب النظري والتطبيقي، مع تقديم مقارنة نوعيّة تهدف إلى تمكين التّواصل بين عالم اللّغة وعالم الحاسوب في سبيل التّهوض بهذا المجال؛ الذي أصبح يُشكّل تحديًا كبيرًا في الحياة العامّة وفي مجال البحث الأكاديمي، كما لا ننسى أنّ له فوائد اقتصاديّة تتنافس فيه كبرى الشركات.

2- معالجة اللّغة الطبيعيّة: Natural Language Processing

تُعرف المعالجة الآليّة أو الحاسوبية للّغة الطبيعيّة (NLP) بأنّها أحد أبرز العلوم البينيّة الحديثة التي تجمّع بين اللغويين وعلماء الحاسوب، وتُعدّ معالجة اللّغة الطبيعيّة مجالاً يسعى إلى إنشاء مفاهيم وإيجاد طرقٍ وبناء برامج قادرة على فهم ودراسة وإنتاج لغاتٍ لتُمكّن تفاعل الإنسان مع الحاسوب من خلال الكتابة والكلام، وبتعبير آخر تُساعد معالجة اللّغة الطبيعيّة (NLP) أجهزة الكمبيوتر على تحديد الطّرق التي يستخدم بها البشر اللّغة.¹ وفي تعريفها يقول عبد الكريم الدخيسي بأنّ: "المعالجة الآليّة للّغة الطبيعيّة مجموعة من التقنيّات الحاسوبية لتحليل وتمثيل نصوص اللّغات الطبيعيّة على مُستوى واحد أو أكثر من التحليل اللّغوي بغرض تحقيق تماثل يُمكن الآلة من إنجاز لغةٍ تُشبه إلى حدّ بعيد اللّغة الطبيعيّة".²

إنّ المعالجة الآليّة للّغات الطبيعيّة تحاول الوصول إلى فهم اللّغة الطبيعيّة وتمثيلها بصورة تُشبه إلى حدّ ما اللّغة الطبيعيّة، وهذا انطلاقاً من النظرة الجديدة للّغة -خاصّة مع

أبحاث تشومسكي وما بعده- والذي عملَ على "بناء آلات ونماذجٍ صوريّة تُحاكي خصائص اللّغة البشريّة"³.

لقد أدّت أبحاث تشومسكي في الاشتغال على صياغة أنحاء صوريّة تصف وتحاكي عمليّتي فهم وإنتاج اللّغة إلى تطوير أنظمة وبرامج في المعالجة الآلية للّغة، وذلك بعدَ تضافر الجهود وحُدوث تلك الصّلة بين اللغويين والحاسوبيين؛ فإذا ما عُدنا إلى استخدام الحواسيب في معالجة اللّغة، وبالضبط في خمسينات القرن الماضي (منذ 1950) وقبل حُدوث هذا التقارب بين علماء اللّغة وعلماء الحاسوب؛ نجد أنّ أولى البدايات كانت عندما تمّت صياغةُ برامجٍ يُمكنها ترجمة مقاطع نصيّة قصيرة من لغة إلى أخرى (كانت غالبًا من الروسيّة أو الألمانية إلى الإنجليزيّة)، فقد تمّ استثمار جُهدٍ كبير في إنشاء قواميس ميكانيكيّة مُفيدة لترجمة النّصوص غير المقيدة، ولكن عندما تمّ وضع هذه القواميس على التّجربة باستخدامها على عيّناتٍ كبيرة من المُدخلات (inputs) اتّضح أنّ التّرجمة أصبحت صعبةً جدًّا، وهنا كان لزامًا صياغةُ برامجٍ جديدة للتعامل مع المشاكل التي لم تأخذ بعين الاعتبار العيّنات النصيّة السّابقة، فأدّت التّعديلات الجديدة إلى تداخلٍ بين البرامج، والتي كان يُنظر إليها سابقًا بأنّها كافية في عمليّة التّرجمة.

وفي خضمّ هذه الأبحاث والجهود من علماء الحاسوب، وبعدَ حوالي عشرة (10) إلى خمس عشرة (15) سنة من العمل؛ اتّضح أنّه من المستحيل الخروج بنتيجة وفق هذه الشّكيلة، فتمّ بهذا التخلّي عن أعمال التّرجمة الآلية التي كانت سائدةً آنذاك والاتّجاه نحو تحليل اللّغة وفهمها.⁴

وبهذا، فإنّ أولى بدايات المعالجة الآلية للّغة كانت عملاً يمكن وصفه بالفردانية بالاعتماد أكثر على المُعطيات والبرامج الحاسوبية دون الاستناد إلى القواعد التي تفرضها اللّغة. وكان ذلك في مجال التّرجمة الآلية التي كانت تعتمد على نظام مقابلة كلمة من اللّغة المصدر بنظيرها في اللّغة الهدف، وذلك بالاستناد إلى القواميس الإلكترونيّة المُجهّزة سابقًا، وإن كانت هذه البرامج ناجحةً في ترجمة الجمل والعبارات البسيطة إلا أنّها فشلت في التّعامل مع الجمل المعقّدة والنّصوص الكبيرة الحجم، ولم تستطع تمثيل المعنى الحقيقي والمقصود للجمل والنّصوص، إضافة إلى عدم القدرة على تحديد المعنى الدّلالي الدّقيق للعديد من الكلمات داخل سياقها، وبهذا عدم حصول ترجمة صحيحة وسليمة؛ الأمر الذي أدّى إلى

حُدوث الاتّصال والتّقارب بين اللّسانيّين والحاسوبيّين بالاستناد إلى نظريّات لغويّة تُتيح الفهم الصّحيح والأمثل للجمل والتّصوص وتحليلها.

كان لتشومسكي في نظريّته التوليدية-الذي حاول عبرها زصد أنحاء صوريّة ذات منطق رياضي- الدور الكبير في صورنة اللّغة، فأسهّم التحو التوليدي "بشكل فعّال في تطوّر الدّراسات الحاسوبية من خلال البحث في خصائص اللّغات الصوريّة ومدى كفايتها في بناء الأوصاف التركيبيّة والدلاليّة للّغات الطبيعيّة، وبذلك استلهم الحاسوبيون حصيلة أبحاث اللّسانيّين لتطوّر خوارزميّات تقبل الإدماج في برمجيات حاسوبية صُمّمت لأجل إنجاز معالجة آليّة للّغة الطبيعيّة، وتوفير المُحلّلات الصرفيّة والتركيبيّة الملائمة لوصف اللّغة الطبيعيّة"⁵.

لقد مثّل تشومسكي-من خلال نظريّته التوليدية- تلك العلاقة بين المنطق الرياضي واللّسانيات، وجسّد اللّغة على أنّها نمطٌ رياضيٌّ؛ سعيًا منه إلى الدّقة والعلميّة التي ما فتأت تُنادي بها اللّسانيات الحديثة، وهذا بالاعتماد على اللّغات الصوريّة؛ "فقد حاولت اللّسانيات الصوريّة صياغة خرائط للّغة، بما يتلاءم مع أدوات الاشتغال البرمجية، وما زالت تخوض في هذا منذ صدور كتاب البنى التركيبيّة لتشومسكي (1957) والتي قُلبت في خضمّها الآليّة المفاهيميّة التحليليّة في اللّسانيات إلى تقنيّات رياضيّة حاسوبية"⁶.

إنّ هذا التّظنّ الجديد في فهم اللّغة وتحليلها جعل اللّسانيات نسفًا مُنفتحًا على مختلف الأنساق المعرفية، كعلم النّفس والرياضيات والمنطق، وأدّى هذا فيما بعد إلى فهم آليّات حُدوث اللّغة واكتسابها، وأيضًا إلى صياغتها في قوالب ونماذج رياضيّة، ليتمّ استثمار هذه الأبحاث في إنشاء برامج واستحداث نظريّات تُمكن الحاسب الآلي من مُعالجة اللّغة الطبيعيّة معالجةً سليمة من خلال فهم اللّغة في البداية ومن ثمّ توليدها.

3- فهم وتوليد اللّغة الطبيعيّة: (NLU/NLG)

تُطالِعُنا في البحث اللّساني الحاسوبي العديد من المُصطلحات الواجب أوّلاً فهمها وتحديدُ الفروق بينها، ومن أهمّها ما تعلق بمعالجة اللّغة الطبيعيّة (NLP) ومُصطلحين آخرين هما من صميم البحث في المعالجة الآليّة للّغة، ويتعلّق الأمر بفهم اللّغة الطبيعيّة Natural Language Understanding وتوليد اللّغة الطبيعيّة Natural Language Generation.

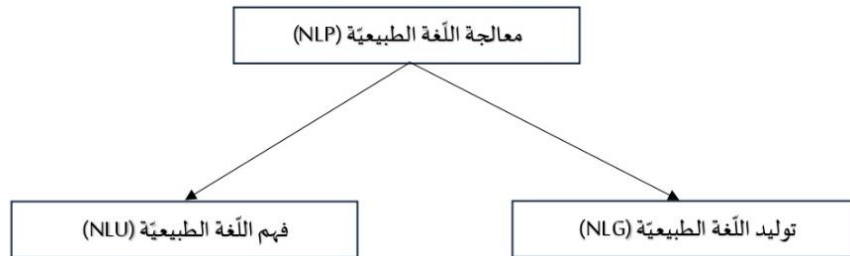
لتوضيح هذا قليلاً، ننتقل أولاً من المعالجة البشرية للغة، والتي تمثل الملكة والكفاية اللتين يستعملهما الإنسان في معالجة اللغة في الذهن، ولتبسيط ذلك نقدّم المثالين الآتيين:

بِدَلَّةُ خَالِدِ الْأَنْبِقِ	بِدِلَّةُ خَالِدِ الْأَنْبِقَةِ
(2)	(1)

في المثالين (1) و (2) يتم معالجة الجملتين وتحليلهما من أجل الوصول إلى دلالتيهما عن طريق مجموعة من القواعد التي تعبر عن مجموعة من الكفايات المُخترَنة في الذهن، والتي تسمح لنا بإدراك الفرق بين الجملتين، خاصة وأنّ الجملتين تتشابهان من حيث البنية السطحية، لكن بتفعيل تلك القواعد المُخترَنة سلفاً يُمكن تحديد العلاقات بين العناصر المُكوّنة لكلّ جملة.

في المثال (1) تمّ إسناد لفظة (الأنبيق) إلى المسند إليه (بِدَلَّةُ خَالِدِ) أمّا في المثال (2) فهناك علاقة تبعية بين لفظة (خالد) ولفظة (الأنبيق)، وبعد تشغيل هذه الإجراءات التحليلية يُمكن الوصول إلى أنّ المثال (1) يريد صاحبه الإخبار بأنّ البدلة أنيقة، أمّا في المثال (2) فالمعنى المقصود هو الإخبار عن أناقة خالد.

هذه الإجراءات التي نمارسها هي التي تتعلق بفهم اللغة، أمّا توليدها فيتعلق بالمرسل الذي يَضَعُ أولاً مجموعة من الاعتبارات النحوية والمقصديّة والسياقية في إنتاج الجملة أو الخطاب، وبهذا فإنّ معالجة اللغة الطبيعيّة تتكوّن من مجموعتين فرعيتين هما: فهم اللغة وتوليد اللغة.



غير أنّ الحديث اليوم هو عن معالجة آية للغة، وهي بالتأكيد تختلف عن تلك العمليات التي قدمناها في المثالين السابقين؛ ذلك أنّ التعامل مع الحاسوب وليس مع العقل البشري، وهذا ما تشتغل عليه اللسانيات الصورية في إيجاد قوالب وأنماط رياضية تمكن من محاكاة اللغة البشرية حاسوبياً.

4- التقسيم وتجزئة الجملة: Tokenization and Sentence Segmentation

يُعدّ التقسيم أو ما يُعرف في الأدبيات الحاسوبية Tokenization من العمليات الأساسية في معالجة اللغات الطبيعية؛ إذ له دور هامّ في تحويل اللغة الطبيعية إلى نسق مفهوم للكمبيوتر يُمكنه من معالجة وفهم اللغة، ففي مجال التحليل اللغوي الحاسوبي لا بُدّ أولاً من تحديد الأجزاء المُشكّلة للكلمات والجمل تحديداً واضحاً.

إنّ تجزئة النصّ أو الجمل مرحلة عادةً ما يتمّ تجاهلها في الحديث عن المعالجة الآلية للغة، لكنّه أمر هامّ ولا يخلو منه أيّ برنامج من البرامج التي تعتمد على معالجة اللغة NLP، نظراً لأنّ الكلمات والجمل المُحددة في هذه المرحلة هي الوحدات الأساسية التي يتمّ تمريرها إلى المعالجة الأخرى مثل: المُحلّلات الصرفية (morphological analyzers)، وُضع علامات على جزء من الكلام (part-of-speech taggers) ونُظم استرجاع المعلومات (information retrieval systems) وغير ذلك.⁷

إنّ التقسيم هو عملية تفكيك تسلسل الأحرف في النصّ عن طريق تحديد حدود الكلمة، والنقاط التي تنتهي فيها كلمة ما وتبدأ أخرى، وفي مجال اللغويات الحاسوبية والمعالجة الآلية يُشار إلى الكلمات التي يتمّ تحديدها باسم التوكن (Token) أو الرمز، ويُعرف أيضاً بتقسيم الكلمات word segmentation، وكثيراً ما يُستخدم هذا المصطلح مُرادفياً لمصطلح Tokenization.

وَنعني بتقسيم الجملة أو تجزئة الجملة Sentence Segmentation: "عملية تحديد الوحدات التي قد تتكوّن من كلمة واحدة أو أكثر، وتتضمّن هذه المهمة تحديد حدود الجملة بين الكلمات في جمل مختلفة، فمُعظّم اللغات المكتوبة تحتوي على علامات ترقيم تنتهي غالباً عند حدود الجملة، ويُشار إلى تجزئة الجملة باكتشاف حدود الجملة Sentence Boundary Disambiguation أو توضيح حدود الجملة Sentence Recognition"⁸، وإنّ جميع هذه المصطلحات تشير إلى كيفية تقسيم النصّ إلى جمل مُواصلة المعالجة الآلية.

وعلى مستوى الجملة الواحدة مثلاً، فإنّ التّفريق بين الكلمات غالباً ما يكون بالبيّاض أو الفراغ بين كلمة وأخرى، وهذا ما يسمّح بتحديد وفصل كلّ كلمة عن أخرى. إنّ هذا الأمر قد يبدو للوهلة الأولى بسيطاً، لكنّه في حقيقة الأمر عمليّة معقّدة وهامّة جدّاً؛ إذ بها يُحدّد مسارُ المعالجة الآليّة ونجاحُ العمليّة التحليليّة للغة، وتمكين الحاسوب من الوصول إلى المعنى، والتّظر في السّياقات المختلفة للجمل والفرق بينها بتحديد الوحدات المتكرّرة في الجمل، والاعتماد أيضاً على علامات التّرقيم، وهذا الأمر لا زال يُشكّل تحديّاً كبيراً في مجال معالجة اللّغة الطّبيعية NLP (حتّى في اللّغة الإنجليزيّة)، فضلاً عن اللّغات التي لا زالت معالجتها حاسوبياً في مراحل متأخّرة نظراً لطبيعة اللّغة ذاتها أو لقلّة الجهود في هذا المجال، كاللّغة العربيّة التي تمتاز بخاصيّة الاشتقاق وتعدّد المعنى باختلاف حركة أو حرف واحد، أو حتّى بسبب علامة من علامات التّرقيم.

ومن التّماذج اللّغويّة الكبيرة (Large Language Model) التي تعتمد على التّرميز وتجزئة الجمل في نظام المعالجة اللّغويّة NLP الخاصّ بها نموذج الدّكاء الاصطناعي: GPT، والذي سنقدّم مثلاً توضيحياً بالتّطبيق على GPT-3,5 & GPT-4 عبر موقع Tokenization الخاصّ بنموذج GPT: (<https://platform.openai.com/tokenizer>)، ولكن قبل ذلك سنطلب منه التّفريق بين الجملتين: "لا عافاك الله" و "لا، عافاك الله" عبر روبوت الدّردشة الخاصّ بالنّموذج: ChatGPT3,5

ChatGPT 3.5

You
الفرق بين لا عافاك الله و لا، عافاك الله

ChatGPT
عبارة "لا عافاك الله" و "لا، عافاك الله" هما تعبيران في اللغة العربيّة يستخدمان بمعاني مختلفة، وقد يتم التّفاوت بينهما حسب السياق الذي يستخدمان فيه. إليك الفرق بينهما:

- "لا عافاك الله":
 - يستخدم هذا التّعبير للتّعبير عن الرّفص أو عدم القبول، ويعني حرفياً "لا تحسن الله إليك".
 - مثال: إذا قال شخص لآخر "هل تستطيع مساعدتي في هذا الأمر؟"، وأجاب الشخص بـ "لا عافاك الله"، يعني ذلك رّفص المساعدة.
- "لا، عافاك الله":
 - هنا يتم استخدام "لا" للرد على سؤال معين أو اقتراح بالرفص، ثم يضاف "عافاك الله" كطريقة لطلب العفو أو السماح.
 - مثال: إذا قال شخص لآخر "هل تذهب معي إلى المتجر؟"، وأجاب الشخص بـ "لا، عافاك الله"، يعني ذلك رّفص الدّعوة بأدب واحترام، وتطلب السماح.

باختصار، الفارق الرئيسي بين العبارتين هو أن "لا عافاك الله" تعبير عن رّفص صريح، بينما "لا، عافاك الله" يُستخدم للرد بالرفص مع تقديم اعتذار أو طلب عفو.

Message ChatGPT..

ChatGPT can make mistakes. Consider checking important information.

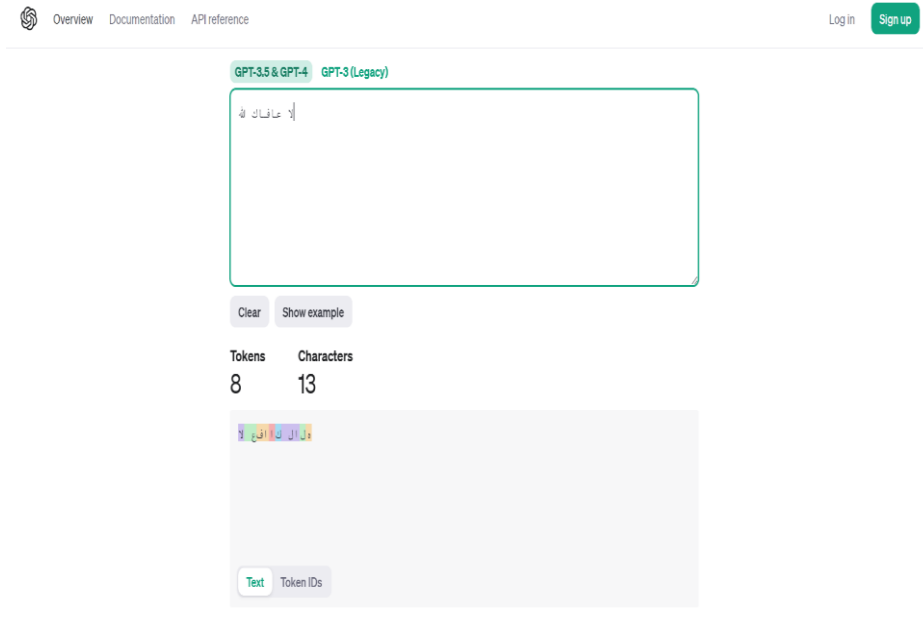
صورة رقم (01): توضّح دردشة عبر ChatGPT3,5

معالجة اللغة الطبيعية: فهم وتوليد اللغة الطبيعية بناء الملائمة أنموذجًا

العدد الثالث عشر / العدد الرابع / ديسمبر 2024

يظهر لنا من الصورة أعلاه أنّ روبوت الذكاء الاصطناعي حاول الوصول إلى فرق بين الجملتين، وأبرز أنّ الاستعمال يختلف بين الجملتين تبعًا للسياق، فالمثال الأول قد يأتي في الدعاء على الشخص بعدم المعافاة، أمّا في الجملة الثانية فهي دعاء بأن يعافى الله الشخص الذي نكلمه، كما أوضح لنا أنّ الفرق بين الجملتين يكمن في الحرف "لا" والذي قد يكون في الأول نفيًا للعافية، بينما في الثاني هو نفي لشيء سابق أو إجابة عن سؤال أو طلب، وبعدها تقديم العفو، وهذان المثالان على بساطتهما إلا أنّهما يمثلان مبحثًا من مباحث البلاغة التقليدية (الفصل والوصل)، وقد لاحظنا كيف حاول الروبوت تحديد المعنى، وأيضا توليد مثال يوضح ذلك.

هذا الظاهر لنا عبر روبوت الدردشة هو نتاج ميكانيزمات وخوارزميات وتخزين مدونات لغوية وترقيمية سلفًا تُمثل قاعدة بيانات، والتي مكّنت الروبوت من الوصول إلى الفرق بين الجملتين المكتوبتين، وإليك الصورتين الآتيتين توضّحان عملية تقسيم الجملتين السابقتين:



صورة رقم (02): عملية تقسيم جملة "لا عافاك الله"

Overview Documentation API reference Log in Sign up

GPT-3.5 & GPT-4 GPT-3 (Legacy)

لا، عافاك الله

Clear Show example

Tokens	Characters
10	15

لا، عافاك الله

Text Token IDs

صورة رقم (03): عملية تقسيم الجملة "لا، عافاك الله"

يظهر لنا من الصورتين (02) و (03) أنّ عملية تقسيم الجملتين كانت تستند إلى مُكوّنات كلّ جملة، ففي الجملة الأولى نجد: 08 Tokens و 13 Characters، في حين في الجملة الثانية نجد: 10 Tokens و 15 Characters، حيث تشير tokens إلى عدد الأحرف المشكّلة للجملة، أمّا Characters إلى عدد الأحرف وكذا الرّموز (في المثالين: الفاصلة والبياض أو الفراغ بين الوحدات والكلمات).

إنّ عملية تقسيم وتجزئة الجمل لها دورها في العملية التحليلية لبناء محلّلات نحويّة وصرفيّة، ومثال ذلك الفرق بين: أغلق الأبواب وغلق الأبواب، فلو بقينا مع روبوت الدردشة نلاحظ أنّه يقوم بتحديد الفرق تبعاً للصيغة الصرفيّة وتحديد مكوّنات كلّ صيغة للوصول إلى المعنى:

ChatGPT 3.5

⌵

You

الفرق بين أغلق الأبواب وغلقت الأبواب

ChatGPT

تعبير "أغلق الأبواب" و"غلق الأبواب" يستخدمان للدلالة على إغلاق الأبواب، ولكن يمكن أن يكون هناك بعض الفروق في الاستخدام والدلالة. إليك الفرق بينهما:

1. "أغلق الأبواب":

- يشير عادةً إلى عملية إغلاق الأبواب بشكل عام.
- "قد يستخدم في اللغة العامية بمعنى "قم بإغلاق الأبواب".

2. "غلق الأبواب":

- يشير أحياناً إلى عملية أكثر حزمًا أو إلى توجيه صريح لإغلاق الأبواب.
- قد يُستخدم في السياقات الأكثر رسمية أو التي تحمل تهديدًا أكبر، حيث يشير إلى توجيه لإغلاق الأبواب بسبب خطر محتمل مثلًا.

بشكل عام، الفرق بينهما يكمن في درجة التحديد والإيضاح، حيث إن "غلق الأبواب" يمكن أن يكون أكثر تحديداً أو توجيهياً مقارنة بـ "أغلق الأبواب" الذي يمكن استخدامه بشكل أكثر عمومية.

⌵ ⌵ ⌵

Message ChatGPT...

⌵

صورة رقم (04): الفرق بين أغلق الأبواب وغلقت الأبواب

ورغم ما تمّ الوصول إليه، غير أنّ الأمر ما زال معقّداً وصعباً خاصّة في النصوص الكبيرة الحجم، وتعدّد السياقات، وكذا إلى طبيعة اللّغة البشريّة في حدّ ذاتها، وأنّ المعنى غالباً ما يخضع لاعتبارات نصية وأخرى مقامية وتداولية.

5- تلخيص النصوص: Text Summarization

من الآليات والبرامج المُستحدثة تلك التي تعتمد تقنية تلخيص النصوص، والتي تقوم أساساً على نظم المعالجة الآلية للغة NLP، وهناك العديد من المواقع التي تقدّم تلخيصاً للنصوص والمقالات اعتماداً على الذكاء الاصطناعي، وقد تمّ إنشاء العديد من الخوارزميات التي تُمكن من إجراء تلخيص ناجح للنصوص والمقالات عبر فهم الجمل وسياقاتها وكذا توليد نصّ يُحافظ على الهيكل العام للنصّ المصدر.

وقد يكون التلخيص من مستند أو نصّ واحد إلى نصّ آخر أقلّ حجماً، أو من مجموعة من المستندات إلى مستند واحد، ولا شكّ أنّ التلخيص التلقائيّ يقدّم لنا خدمات مهمّة أبرزها اختصار الوقت في الوصول إلى المعنى أو الموضوع الذي بين أيدينا، كما أنّه في الوقت ذاته

يُمكننا من الوصول إلى تحديد الفجوة البحثية لمجموعة من الدراسات والأبحاث بشكل أسهل.

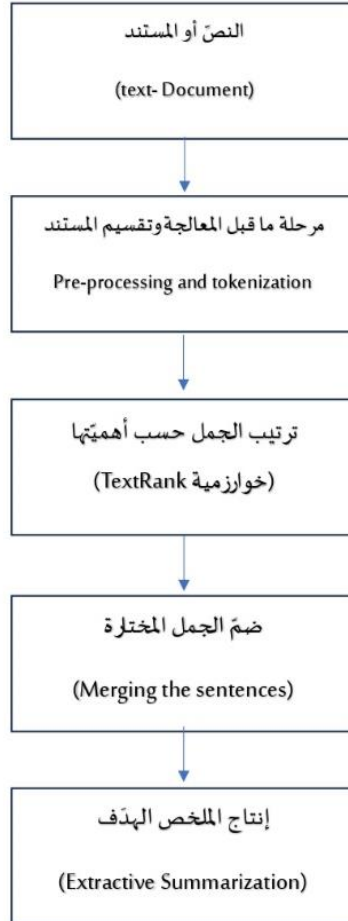
وعليه؛ فالحديث هنا يكون عن نوعين من أنواع التلخيص: الأول يُمكن تسميته بالتلخيص الاستخراجي، وهو ما يعرف بمصطلح: Extractive Summarization والآخر هو تلخيص يُمكن القول بأنه تلخيص إجرائي: Abstractive Summarization .

1.5- التلخيص الاستخراجي: Extractive Summarization

تعتمد تقنية أو أداة التلخيص الاستخراجي في المعالجة الآلية للغة الطبيعية على اختيار أهمّ الجمل في المُستند للحصول على نسخة مُختصرة من النصّ الأصلي⁹، ففي هذا أشبه بعملية تقليص للنصّ عن طريق حذف الجمل الثانوية والتكرارات والشروحات والأمثلة، دون التّغيير في تركيب الجمل والوحدات المُشكّلة لها.

تستند هذه التقنية إلى خوارزمية تمّ إنشاؤها لهذه الغاية وتُعرف بـ TextRank algorithm والتي تقوم على تحديد أهمّ الكلمات الرئيسة في المُستند، وقد تمّ إنشاؤها من طرف الباحثين: ميهالسيا (Mihalcea) و بول تارو (Paul Tarau) وهذا النهج من التلخيص بالاعتماد على هذه الخوارزمية يُمكن من بناء ملخصات دون الحاجة إلى تدريب مسبق، كما يُمكن تطبيق هذه الخوارزمية على العديد من اللغات.¹⁰

وهنا يتّضح لنا جلياً دور الترميز وتقسيم النصّ Tokenization، إذ يُستخدم في تحديد نهاية الكلمات والجمل وحدودها، كما يُمكن من تحديد الكلمات والجمل الأكثر تردداً وبالتالي التوصل إلى أهميتها، وهذا ما قلنا عنه في العنصر السابق أن Tokenization يمثل مرحلة ما قبل المعالجة أو التحويل.



صورة رقم (05): مخطّط توضيحي لعملية تلخيص استخراجي Extractive Summary (إنجاز الباحث)

2.5- التلخيص الإجمالي: Abstractive Summarization

تقوم هذه المهمة على نظم وخوارزميات المعالجة اللغوية الموضوعية لأجل ذلك بعملية أكثر نوعية؛ إذ أنها لا تكتفي بتمييز الجمل المهمة في النص أو المستند الأصلي، بل تقوم بتوليد مخطّط تفصيلي بالمخرجات المطلوبة من كلمات وجمل مختلفة عن جمل المستند الأصلي¹¹؛ فهي تقوم بتحديد الفقرات والجمل والكلمات، ثم الوصول إلى التسيج العلائقي الذي يربط

بينها لفهم المعنى والدلالات فهما يُمكنها من توليد نص آخر أقل حجماً بكثير ومخالفاً من حيث الكلمات والجمل وكذا ترتيبها، لكنّه محافظ على المعنى الأساس للنصّ المصدر. يهدف التلخيص الإجرائي Abstractive Summarization إلى إعادة صياغة النص من أجل إعداد ملخصات أكثر اتساقاً، وهي تعتمد في ذلك على نظام الشبكات العصبية الاصطناعية (Neural Abstractive Text Summarization)، وهو يقوم على تحديد العلاقات الدلالية بين الجمل وال فقرات، وذلك باستخدام التمثيل الدلالي للمستند في توليد اللغة الطبيعية (NLG)، وذلك بتحديد دلالات الأسماء والأفعال والحروف، ثم توليد ملخص بالاعتماد على نموذج: التتابع إلى التتابع Sequence to Sequence model، وهذا النموذج قد تمّ استخدامه بداية في عمليات الترجمة الآلية (Machine Translation) والتعرّف الآلي على الكلام (Speech Recognition)، ويتمثل نموذج التتابع إلى التتابع Sequence to Sequence model في التلخيص الإجرائي (Abstractive Summarization) برسم خريطة لتتابع أو تسلسل مجموعة من المدخلات (كلمات وجمل) للمستند المصدر إلى تتابع مستهدف من الكلمات لإنتاج ملخص¹².

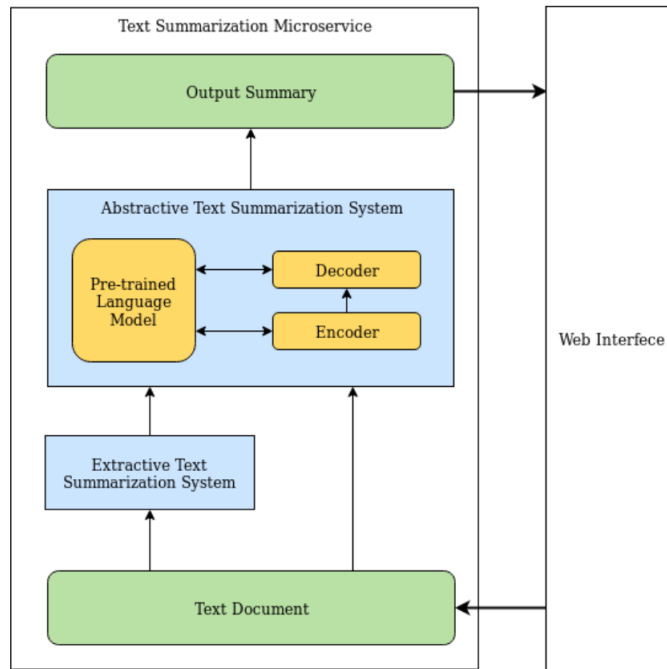
إنّ هذا النموذج القائم على الشبكات العصبية يتعامل مع النصّ المصدر على أنّه مجموعة من الوحدات المتتابعة يتمّ تحويلها إلى وحدات أخرى متتابعة، وهو يتعامل بهذا فقط مع الوحدات التي تشكّل زمناً معيناً، كما هو الحال في النصوص والكلام، وبهذا فهو مساعد في عملية التلخيص الآلي، وقد استعمل أول الأمر في مجال الترجمة الآلية، ونقدّم المخطّط الآتي لتبسيط الأمر:



صورة رقم (06): مخطّط تمثيلي لنموذج التتابع إلى التتابع Sequence to Sequence model (من إنجاز الباحث)

يقدم المخطّط مثلاً عن Sequence to Sequence model في مجال التّرجمة الآليّة، فهذا النموذج يُعدّ النصّ الأصليّ مجموعة من الوحدات المتتابعة تشكّلها الكلمات والحروف، والنصّ الهدف هو أيضاً مجموعة من الوحدات المتتابعة في اللّغة الهدف، والأمر نفسه ينطبق على التلخيص، فالنص أو المستند هو في الأصل مجموعة من الجمل المتتابعة نريد القيام بعملية تلخيص لها في مستند أو نص هدف، بحيث يقدم لنا في التلخيص الإجرائي Abstractive Summarization نصاً آخر محافظاً على المعنى والموضوع، ولكن بلغة وجمل غير تلك الموجودة في المستند المصدر، وذلك بالاعتماد على نظام التوليد (NLG).

ومن ثمّ الاستناد على نموذج دلاليّ باعتماد خوارزميات معيّنة، لتمثيل العلاقة بين المفاهيم والدلالات، باتّباع المراحل السّابقة من تقسيم وترميز الوحدات (الجمل والكلمات) Tokenization، وربّما الاعتماد على معطيات التلخيص الاستخراجي في تحديد الجمل والعبارات المهمّة، وكذا الكلمات المفتاحيّة، ومن ثمّ توليد الملخص: Generate the Summary، والمخطّط التالي يوضّح ذلك:



صورة رقم (07): مراحل بناء ملخص إجرائي تلقائي (Abstractive Summarization)

عند إدخال المستند (Text Document) إلى الموقع أو التطبيق، تتم عمليتان: الأولى عملية تلخيص استخراجي (Extractive Text Summarization System) والثانية هي عملية تشفير وفك التشفير (Encoder/Decoder) حيث يتم تشفير الكلمات والجمل إلى لغة معينة (Paython مثلا) لتحويلها إلى لغة الآلة، ثم التعامل مع نموذج لغوي مدرب مسبقًا (Pre-trained Language Model)، لتشكل المرحلتان الأولى والثانية عملية تلخيص النص الإجمالي عبر نظام المعالجة الآلية، وتم إخراجها إلى نص ملخص (Output Summary).

ورغم كل ما تم ذكره، تبقى فعالية الأمر نسبية، فلا يزال البحث في إيجاد خوارزميات وبناء برامج تلخيص قادرة على إجراء تلخيص سليم للنصوص -خاصة النصوص الكبيرة الحجم- حتى مع البرامج التي تقوم على الذكاء الاصطناعي، فكل هذا يعتمد أساسًا على تحليل صحيح للغة البشرية، والوصول إلى حل ما يعترضها من تغيرات دلالية واعتبارات سياقية وتداولية تتحكم في إنتاج اللغ؛ وهنا يبقى البحث جاريًا في سبيل تمكين الحاسوب والآلة عمومًا من فهم اللغة وتوليدها، وما تلخيص النصوص إلى مثال عن مسار عملية فهم وتوليد اللغة (NLU/NLG) في معالجة اللغة الطبيعية NLP.

6- خاتمة:

ختامًا، إن معالجة اللغة الطبيعية من المجالات التي تمثل تحديًا كبيرًا؛ إذ لا تزال البحوث والجهود الساعية في عمل محاكاة للغة البشرية ومعالجتها آليًا قائمة، وهذا عن طريق تزويد الحاسوب أو الآلة بتلك الآليات المفاهيمية وبناء خوارزميات كفيلة بفهم وتوليد اللغة حاسوبيًا، ويبقى المنطلق في ذلك إيجاد النظرية اللغوية الممكن صورتهما وصياغتهما وفق أنماط رياضية تسمح للحاسوب من فهم المفردات والتراكيب ودلالاتها وتوليفها تلقائيًا، ومن الجهود التي رأيناها في البحث ما تعلق ببناء برامج تلخيص فعالة، والتي لا يزال البحث قائمًا فيها بعد أن أثبتت البرامج الحالية عدم فعالية تامة.

ويمكن تلخيص نتائج البحث في النقاط الآتية:

- معالجة اللغة الطبيعية من الدراسات البينية الحديثة التي أصبحت تجمع بين علماء اللغة وعلماء الحاسوب.

- بناء البرامج القائمة على المعالجة اللغوية وصياغة الخوارزميات يكون أولًا بتحليل

اللغة واكتشاف مظاهرها وخصائصها.

- إيجاد الصّياغة الرياضية المناسبة للّغة هو الكفيل بتجاوز الصّعوبات في معالجة اللّغة الطّبيعية.

- يعدّ التّرميز وتقسيم الجمل والكلمات من الآليات التي تعتمد عليها معالجة اللّغة الطّبيعية بشقّها (الفهم والتّوليد).

- بناء الملخّصات بنوعها (الاستخراجية والإجرائية) يعتمد على نُظم فهم اللّغة وتوليد اللّغة الطّبيعية، وهو من البحوث التي لا تزال الجهود قائمة فيها.

7. مراجع البحث وإحالاته:

- 1 - Souad Larabi and al, Arabic Natural Language Processing and Machine Learning-based Systems, IEEE Access, Vol: 20, 2017, p: 01.
- 2 - عبد الكريم الدخيسي، اللسانيات الحاسوبية والمعالجة الآلية للغة الطّبيعية، عالم الكتب الحديث، الأردن، ط1، 2020، ص25.
- 3 - امحمد سعيد الملاخ، حافظ إسماعيل علوي، اللسانيات التوليدية من نظرية العمل والرّبط إلى البرنامج الأذنوي، دار كنوز المعرفة، عمان، الأردن، ط1، 2016، ص15.
- 4- Gerard Salton & Michael J. McGill, Introduction to modern Information Retrieval, Mc Graw-Hill, Library og Congress Cataloging in Publication Data, 1983, p: 261,262.
- 5 - امحمد سعيد الملاخ، حافظ إسماعيلي علوي، اللسانيات التوليدية من نظرية العمل والرّبط إلى البرنامج الأذنوي، ص26.
- 6- دكيكي عبد الواحد، الدّخلة البسيطة والدّخلة المركّبة من منظور لساني حاسوبي، استكتاب جماعي: اللسانيات الحاسوبية واللغة العربية إشكالات وحلول، إعداد وتنسيق: عمر مهديوي، دار كنور المعرفة للنشر والتوزيع، عمان، الأردن، ط1، 2018، ص176.
- 7- David D. Palmer, Tokenization and sentence: <https://tm-town-nlp-resources.s3.amazonaws.com/ch2.pdf> p:01.
- 8- David D. Palmer, Tokenization and sentence, p:01.
- 9- Federico Barrios & al, Variations of the Similarity Function of TextRank for Automated Summarization, arXiv: 16.020306v1, 11 Feb 2016, p: 01.
- 10- p :02.
- 11- G. Vijay Kumar & al, Text Summarization Using NLP, IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial, 2021, p: 61.

12- Ramesh Nallapati & al, Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pages 280–290, Berlin, Germany, August 7-12, 2016, p: 280.

8- مراجع البحث:

1. عبد الكريم الدخيسي، اللسانيات الحاسوبية والمعالجة الآلية للغة الطبيعية، عالم الكتب الحديث، الأردن، ط1، 2020
2. امحمد سعيد الملاخ، حافظ إسماعيل علوي، اللسانيات التوليدية من نظرية العمل والربط إلى البرنامج الأدنوي، دار كنوز المعرفة، عمان، الأردن، ط1، 2016
3. دكيكي عبد الواحد، الدخلة البسيطة والدخلة المركبة من منظور لساني حاسوبي، استكتاب جماعي: اللسانيات الحاسوبية واللغة العربية إشكالات وحلول، إعداد وتنسيق: عمر مهديوي، دار كنوز المعرفة للنشر والتوزيع، عمان، الأردن، ط1، 2018
4. Ramesh Nallapati & al, Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pages 280–290, Berlin, Germany, August 7-12, 2016,
5. G. Vijay Kumar & al, Text Summarization Using NLP, IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial
6. Federico Barrios & al, Variations of the Similarity Function of TextRank for Automated Summarization, arXiv: 16.020306v1, 11 Feb 2016
7. David D. Palmer, Tokenization and sentence: <https://tm-town-nlp-resources.s3.amazonaws.com/ch2.pdf>
8. Gerard Salton & Michael J. McGill, Introduction to modern Information Retrieval, Mc Graw-Hill, Library og Congress Cataloging in Publication Data, 1983
9. Souad Larabi and al, Arabic Natural Language Processing and Machine Learning-based Systems, IEEE Access, Vol: 20, 2017