République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Saad Dahleb Blida -1-
Faculté des Sciences
Département d'informatique

MÉMOIRE DE MASTER II

Spécialité : Ingénierie des Logiciels

THÈME

Analyse des dépendances et apprentissage en profondeur pour la génération automatique de questions

Présenté par :
- **Nehari Manel**
- **Zouaoua Yasmine**

Encadré par :
Mme Ouahrani Leila

Soutenu : 09/07/2023
Devant le jury Composé de
Mme. CHERIGUENE Présidente
Mme. BOUCETTA Examinatrice

Année Universitaire : 2022/2023

Remerciements

Nous exprimons tout d'abord notre gratitude envers Dieu pour nous avoir accordé la force et la santé nécessaires pour mener à bien ce mémoire.

Nous souhaitons sincèrement remercier notre promotrice, Mme Ouahrani Leila, pour son encadrement professionnel, sa présence constante, ses précieux conseils et ses remarques éclairantes. Sans son soutien inestimable, ce travail n'aurait pas pu voir le jour et ne serait pas aussi enrichissant.

Nous tenons également à exprimer notre profonde reconnaissance envers les membres du jury qui ont consacré leur temps à la correction de notre mémoire. Leurs précieuses contributions et leurs suggestions constructives ont grandement contribué à l'amélioration de notre travail.

Nous adressons nos sincères remerciements à nos chers parents, à nos frères et sœurs pour leur amour inconditionnel et leur soutien indéfectible tout au long de notre parcours académique. Leur présence et leurs encouragements ont été une source de motivation constante. Que Dieu les préserve pour nous.

Nous adressons à la fin nos remerciements à nos chers amis et à toute personne qui a contribué de près ou de loin à la réussite de ce travail.

Résumé

La génération automatique de questions à choix multiples est un domaine de recherche prometteur qui vise à faciliter la création de matériel pédagogique, de tests et d'évaluations en générant automatiquement des questions et des options de réponse à partir de textes.

Elle présente de nombreux avantages, notamment la réduction du temps et des efforts nécessaires pour créer des questions.

Notre modèle repose sur une approche linguistique en effectuant l'analyse syntaxique d'un texte arabe, qui fournit des relations des dépendances structurelles, des parties du discours et la reconnaissance des entités nommées. En combinant ces derniers avec la notion des expressions régulières nous sommes optes d'obtenir et de générer des modèles de question après avoir pu sélectionner les mots clés pertinents dans le texte. Par la suite, nous utilisons les modèles des words embedding pour pouvoir générer des distracteurs pour chaque mot clé sélectionné.

Mots clés : la génération automatique de question, les dépendances structurelles,

la génération de distracteurs, la Reconnaissance des entités nommées.

ABSTRACT

Automatic generation of multiple-choice questions is a promising research area that aims to facilitate the creation of educational materials, tests, and assessments by automatically generating questions and answer options from texts. It offers numerous advantages, including reducing the time and effort required to create questions.

Our model is based on a linguistic approach that performs syntactic analysis of Arabic texts, providing information on structural dependencies, parts of speech, and named entity recognition. By combining these elements with regular expressions, we are able to obtain and generate question templates after selecting relevant keywords from the text. Subsequently, we use word embedding models to generate distractors for each selected keyword.

Keywords: automatic question generation, structural dependencies, distractor generation, named entity recognition.

الملخص

يعد التوليد التلقائي الأسئلة متعددة الاختيارات مجالًا بحثيًا واعدًا يهدف إلى تسهيل إنشاء المواد التعليمية و الاختبارات والتقييمات عن طريق التوليد التلقائي للأسئلة وخيارات الإجابة من النصوص.

لأسئلة متعددة الاختيارات فوائد عديدة ، بما في ذلك تقليل الوقت والجهد اللازمين لإنشاء الأسئلة .

يعتمد نموذجنا على نهج لغوي من خلال إجراء تحليل نحوي لنص عربي ، والذي يوفر علاقات تبعية هيكلية وأجزاء من الكلام والتعرف على الكيانات المسماة. من خلال الجمع بين هذه مع مفهوم التعبيرات العادية ، فإننا نختار الحصول على نماذج الأسئلة وإنشاءها بعد أن نكون قادرين على تحديد الكلمات الرئيسية ذات الصلة في النص. بعد ذلك ، نستخدم نماذج تضمين الكلمة لنكون قادرين على توليد مشتتات لكل كلمة رئيسية محددة.

.

الكلمات الرئيسية: إنشاء الأسئلة تلقائيًا ، التبعيات الهيكلية ، توليد المشتتات ، التعرف على الكيان المحدد.

Table des Matières

Tables des figures

Liste des tableaux

Glossaire de termes

Abréviation	Signification
QCM	Questions à choix multiples (Multiple Choice Questions)
GAQ	Génération automatique des questions
GQCM	Génération des questions à choix multiples
NER	Reconnaissance des entités nommées (Named Entity Recognition)
POS	Partie de discours (Part Of Speech)
SRL	Étiquetage sémantique de rôle (Semantic Role Labeling)
TAL	Traitement automatique de langage naturel
BOW	Sac de mots (Bag Of Words)
Glove	Vecteurs globaux pour la représentation des mots (Global Vectors For Word Representation)
Regex	Les expressions régulière (Regular expressions)
GPU	Processeur graphique (Graphics Processing Unit)
API	interface de programmation d'application (application programming interface)

Introduction Générale

L'apprentissage en ligne présente de nombreuses opportunités pour enrichir le processus éducatif traditionnel. Cependant, la question de l'évaluation des connaissances demeure un élément essentiel dans l'apprentissage en ligne et reste un défi sur le plan pédagogique.

Il est indéniable que les experts humains, notamment les enseignants, jouent un rôle irremplaçable dans la vérification des connaissances des apprenants. Bien que divers systèmes de soutien à l'éducation existent, la génération de questions demeure un domaine qui n'est pas suffisamment exploré.

Pourtant, poser des questions revêt une importance cruciale dans le processus d'apprentissage. Cela stimule l'engagement des apprenants, favorise les échanges d'idées, encourage l'innovation et contribue à l'amélioration des performances.

En effet, il existe trois types de questions : les questions ouvertes, qui permettent de recueillir des informations détaillées en posant des questions telles que "pourquoi ?", "comment ?", et "qu'est-ce que ?"; les questions fermées, qui visent à obtenir des réponses précises avec des options telles que "oui", "non" ou "peut-être"; et enfin les questions à choix multiple, qui orientent les apprenants vers un ensemble restreint de réponses possibles.

Les questions à choix multiple offrent de nombreux avantages, tels qu'une évaluation rapide, une méthodologie de notation claire et une économie de temps lors des tests [1]. C'est pourquoi de nombreux concours et évaluations utilisent ce format pour évaluer les compétences des candidats.

La génération de questions et la formulation adéquate de celles-ci sont des éléments clés des technologies d'apprentissage avancées, comme les systèmes de tutorat intelligent, les systèmes d'aide à la recherche d'informations et les systèmes de dialogue homme-machine.

Ces dernières années, les progrès réalisés dans le domaine du traitement automatique du langage naturel ont largement contribué aux applications éducatives, notamment avec la génération automatique de questions.

La génération automatique de questions à partir d'un texte implique la transformation automatique de phrases déclaratives en phrases interrogatives. Cela nécessite l'utilisation de ressources et d'outils tels que la détection d'entités nommées, l'analyse syntaxique et la simplification de phrases. Les applications de cette technologie sont variées, allant de la création de tests et de questionnaires à choix multiples pour faciliter l'apprentissage, aux systèmes de dialogue homme-machine et aux interfaces de questions-réponses interactives.

Dans le cadre de notre travail, nous nous concentrons sur le développement d'un générateur automatique de questions en langue arabe. Contrairement au domaine anglophone, où de nombreuses avancées ont été réalisées, il existe peu de travaux équivalents pour la langue arabe en raison de ses spécificités grammaticales et du manque de ressources et d'outils disponibles. C'est pourquoi nous avons choisi de nous orienter vers un système d'interrogation à choix multiple.

Objectifs

L'objectif principal de ce travail est de construire un générateur automatique des questions à choix multiples en arabe dont il fournit des questions avec plusieurs choix distracteurs.
Nous abordons le travail avec une approche linguistique qui se base sur les dépendances structurelles (dependency parsing, POS Tagging, ...) pour générer à partir d'une phrase (l'item de la question), les concepts clés sur la base desquels les transformations sont faites pour générer automatiquement des questions à choix multiples. Les distracteurs de réponses sont construits automatiquement en utilisant le calcul de similarités à base de cooccurrences incorporés dans des words embeddings (issus d'un pré-entraînement profond). La génération automatique doit passer par la réalisation des étapes suivantes :

- Explorer les approches de génération automatique de questions.
- Explorer les modèles d'apprentissage profond de Word Embeddings pré-entrainés pour les utiliser dans le calcul des distracteurs de réponses.
- Explorer les outils et ressources (Parsing, Pos tagging, ...).
- Concevoir et implémenter le modèle de génération de questions.
- Évaluer la qualité du modèle de génération des questions .

Structure de mémoire :

Afin de réaliser notre travail, notre mémoire est structuré en 5 chapitres différents :

- Chapitre 1 : nous présentons un état de l'art sur la génération automatique des questions à choix multiple .
- Chapitre 2: ce chapitre est consacré à la conception du modèle de génération de questions à choix multiple.
- Chapitre 3: porte sur la réalisation et l'évaluation de notre système élaboré.

Chapitre 1: Etat de l'art

1. Introduction

La construction manuelle des questions est un processus complexe qui nécessite une expérience et des ressources, elle prend beaucoup du temps et énergie ce qui entrave et ralentit l'utilisation des activités éducatives, par exemple fournir des questions des tests ou bien des examens qui nécessite un large éventail de question pour les enseignants, satisfaire la curiosité et parfaire les connaissances et combler les lacunes des étudiants. Pour éviter les problèmes reliés à la construction manuelle des questions et pour satisfaire le besoin d'un approvisionnement continue des questions, des techniques de génération de question automatique ont été introduites, alor nous présentons dans ce chapitre le concept de génération des question automatiques à choix multiple, les différentes approches utilisés dans ce concept, son domaine d'application et enfin nous allons présenter les travaux réalisés dans le contexte de la langue arabe connexes les plus récents.

2. La génération automatique des questions (GAQ)

Le système de génération de questions peut générer automatiquement des questions à partir du texte donné. C'est le processus qui consiste à prendre du texte en entrée et à générer des questions en sortie. Les questions générées peuvent prendre différentes formes, y compris des questions à choix multiples (QCM), des questions ouvertes, des questions à réponse courte etc. Le but de la GAQ est d'automatiser le processus de création de questions pour améliorer l'efficacité et la qualité des questions utilisées dans l'apprentissage et l'évaluation [1]. La GAQ est utilisée dans des domaines tels que l'éducation, la formation, l'évaluation des compétences et la recherche, etc .

3. La génération des questions à choix multiples (GQCM)

Les avancées technologiques dans le domaine de l'intelligence artificielle ont transformé de nombreux aspects de notre vie quotidienne, y compris l'apprentissage en ligne. Alors que les plateformes d'apprentissage en ligne gagnent en popularité, l'évaluation des connaissances reste un défi majeur. Les enseignants sont essentiels pour vérifier les connaissances des apprenants, mais le besoin de systèmes de soutien

à l'éducation automatisés est en augmentation. Les systèmes de génération automatique de questions à choix multiple peuvent jouer un rôle crucial en créant des questions pertinentes, de qualité et en grand nombre pour les examens, les évaluations et les quiz. Outre leur utilisation dans l'éducation, ces systèmes peuvent améliorer les interactions homme-machine dans de nombreux autres domaines.

Lorsqu'on crée un QCM, il est essentiel de comprendre les trois éléments principaux qui le composent :

- **L'énoncé ou la question :** C'est la partie de la question qui pose un problème ou une demande de réponse .
- **Le mot clé :** C'est la ou les réponses que le participant doit identifier comme étant correcte(s) .
- **Les distracteurs :** Ce sont les options de réponse qui sont incorrectes et qui sont conçues pour tromper et distraire le participant . [1]

Par exemple, on considère le QCM suivant [1] :

ما هو أكبر كوكب في نظامنا الشمسي ؟

1)الأرض

2) المشتري

3) الزهرة

4) نبتون

- L'énoncé de notre QCM est : ما هو أكبر كوكب في نظامنا الشمسي ؟
- Le mot clé est : المشتري
- Les distracteurs sont : الأرض , الزهرة, نبتون

4. Processus de construction d'un QCM

La création d'un questionnaire à choix multiples (QCM) peut sembler simple, mais c'est en réalité une tâche complexe qui nécessite plusieurs étapes clés pour garantir la qualité de la question posée. En effet, un QCM bien conçu doit être pertinent, clair, non ambigu, et permettre de mesurer efficacement la compréhension et la maîtrise d'un sujet par le répondant. La figure 1.1 ci-dessous présente un récapitulatif des différentes étapes à suivre pour créer un questionnaire à choix multiples (QCM).

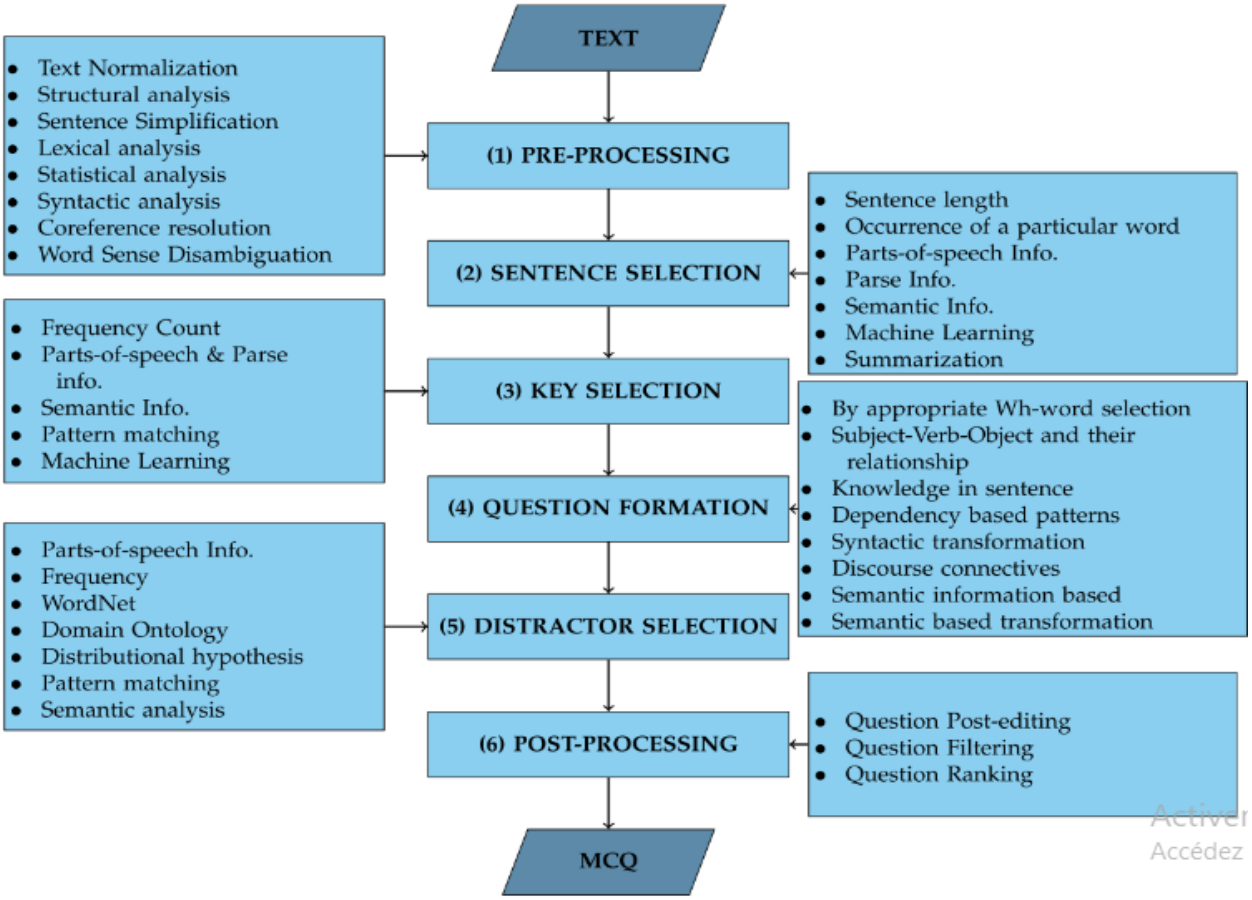

Figure 1.1 : Schéma expliquant les approches utilisées dans la génération d'un QCM. [1]

Dans les paragraphes suivants, nous allons expliquer en détail les différentes étapes présentées dans la figure 1

4.1 Prétraitement du texte entré : Le prétraitement de texte est une étape importante dans le traitement automatique du langage naturel (TALN) qui consiste à nettoyer et à transformer le texte brut afin de le rendre plus facilement exploitable par les algorithmes informatiques. Nous citons quelques uns :

4.1.1 Normalisation du texte : La normalisation fait référence à une conversion du texte d'entrée dans le format requis et suppression du contenu inutile du texte. [1]

4.1.2 Analyse lexicale : Cela implique de diviser le texte du document en une série de mots, de symboles et de nombres. De plus, les inflexions peuvent affecter divers modules individuels de la tâche. Par conséquent, la racinisation qui trouve la forme racine du mot.[1]

4.1.3 Analyse syntaxique : Différents niveaux d'analyse syntaxique y compris le marquage des parties du discours (attribuant des parties de parole de chaque mot),

reconnaissance d'entité nommée (trouver noms de catégories prédéfinies d'un texte), analyse des dépendances (génération de la structure d'analyse de la phrase) . [1]

4.2 La sélection de la phrase : Chaque phrase d'un texte n'est pas en mesure de générer une question valide. Seules les phrases qui contiennent un fait discutable peuvent servir de candidats pour générer des QCM. Par conséquent, la sélection des phrases joue un rôle majeur dans la tâche de génération automatique de QCM. Plusieurs approches ont été utilisées dans la littérature pour sélectionner des phrases informatives et discutables à partir d'un texte. Nous présentons ces deux dernières:

4.2.1 Analyse des dépendances structurelles : Le terme "Analyse de Dépendance" (AD) fait référence au processus d'examiner les dépendances entre les mots d'une phrase afin de déterminer sa structure grammaticale. Une phrase est divisée en plusieurs sections principalement sur cette base. Le processus repose sur l'hypothèse qu'il existe une relation directe entre chaque unité linguistique dans une phrase. Ces liens sont appelés dépendances.[2]

Exemple : محمد يقرأ الكتاب في المكتبة

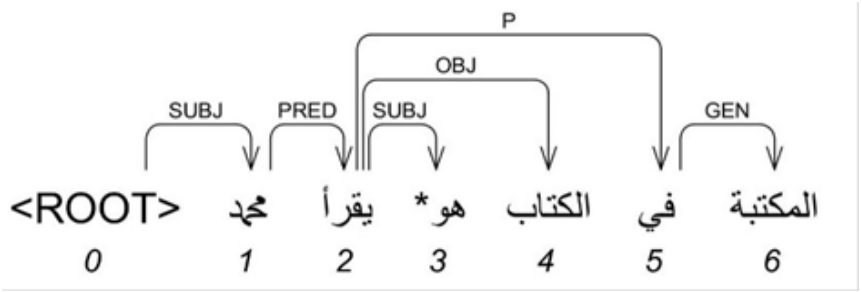

Figure 1.2 : Exemple de dépendance structurelle d'une phrase en arabe [3]

4.2.2 Partie du discours (part-of–speech) : La partie du discours se réfère à la classification grammaticale des mots en fonction de leur rôle et de leur fonction dans une phrase. Chaque mot est classé en une partie du discours particulière, en fonction de sa structure, de sa fonction grammaticale et de son sens. Les parties du discours en arabe incluent notamment le nom, le verbe, l'adjectif, l'adverbe, la préposition, la conjonction et la particule. La connaissance des parties du discours en arabe est essentielle pour la compréhension de la structure grammaticale de la langue arabe et pour la création de phrases correctes et cohérentes. [4]

Il existe plusieurs outils de traitement automatique de la langue arabe qui permettent d'effectuer l'analyse de parties du discours. Voici quelques-uns des outils les plus couramment utilisés :

- **Stanford Arabic Parser :** un analyseur de syntaxe et de partie de discours (POS) pour l'arabe, développé par l'université de Stanford.[5]

- **MADAMIRA :** un outil open-source d'analyse morphologique et de parties de discours pour l'arabe, développé par l'université de Columbia. [6]
- **Farasa :** un outil open-source de traitement de la langue arabe qui inclut l'analyse morphologique et de parties de discours pour la langue arabe, développé par l'université King Abdulaziz. [7]
- **Stanza:**Stanza est une collection d'outils précis et efficaces pour l'analyse linguistique de nombreuses langues humaines. Du texte brut à l'analyse syntaxique et à la reconnaissance d'entités, Stanza apporte des modèles NLP de pointe aux langues de votre choix. [8]

4.2.3 Reconnaissance d'entité nommée (NER) : La reconnaissance d'entité nommée (NER) fait référence à une tâche d'extraction de données qui est chargée de rechercher, de stocker et de trier le contenu textuel dans des catégories par défaut telles que les noms de personnes, d'organisations, de lieux, d'expressions de temps, de quantités, de valeurs monétaires et de pourcentages. La reconnaissance des entités nommées est un système d'intelligence de pointe qui fonctionne avec presque l'efficacité d'un cerveau humain. Le système est structuré de telle manière qu'il est capable de trouver des éléments d'entité à partir de données brutes et peut déterminer la catégorie à laquelle appartient l'élément. Le système lit la phrase et met en évidence les éléments d'entité importants dans le texte.

Exemple : "جون سميث هو خبير في المعلوماتية", le système NER va extraire "جون سميث" en tant qu'entité personne et "المعلوماتية" en tant qu'entité de champ, et va les étiqueter en conséquence. [9]

4.3 La sélection de mot clé : Il est évident que tous les mots d'une phrase informative ne peuvent pas servir de mot clé. Par conséquent, la sélection de mot clé est une étape essentielle qui détermine le mot, le n-gramme ou l'item dans la phrase sélectionnée qui sera effacée [1]. Nous discutons ci-dessous des approches utilisées dans la littérature de notre contexte:

4.3.1 Nombre de fréquences : Le nombre de fréquences des mots a été utilisé comme critère de sélection dans un certain nombre de systèmes QCM. Cependant, seuls le comptage de fréquence ne fournit pas d'indice suffisant. Par conséquent, certaines informations supplémentaires sont également utilisées avec la fréquence pour faire la sélection finale. Certains fois, Tf * Idf a été utilisé à la place du simple terme fréquence. [1]

4.3.2 Informations sur la partie du discours et l'analyse : il est observé qu'une partie particulière du discours ou une catégorie d'analyse devient dominante en tant que mot-clé potentiel dans certains domaines ou applications spécifiques. [1]

4.4 Génération de la question : Après la sélection de la clé, la tâche suivante devient une transformation de la phrase déclarative à la forme interrogative. Il existe plusieurs approches pour la génération automatique de questions. En voici quelques-unes :

4.4.1 Approche basée sur les dépendances structurelles :

L'approche de dépendances structurelles est une méthode d'analyse linguistique qui se concentre sur la relation de dépendance entre les mots d'une phrase. Cette approche considère chaque mot comme un nœud et les relations de dépendance entre les mots comme des arcs qui les relient. Les relations de dépendance peuvent inclure des informations telles que les relations sujet-verbe, les relations objet-verbe.

Afzal et Mitkov [10], ont utilisé des modèles basés sur les dépendances pour la formation de questions. À partir de l'arbre de dépendances de la phrase, ils ont identifié le verbe principal et la portion qui sera posée comme question. Ensuite, ils ont retiré les parties inutiles et sélectionné le mot interrogatif approprié pour transformer la phrase en une question.

- **Approche basée sur les règles :** Cette approche utilise des règles linguistiques pour générer des questions en fonction de certaines propriétés du texte source. Elle varie de système à système [1]. Ceux-ci sont résumés ci-dessous:
  - ➢ **Transformation syntaxique**: **Heilman et al** [11] ont identifié les phrases de réponse et ont effectué une transformation syntaxique pour générer des questions. Leurs principales étapes comprennent : le marquage des phrases non déplaçables, la génération de phrases de question possibles en utilisant une approche basée sur des règles, la décomposition du verbe principal, l'inversion sujet-auxiliaire, la suppression des réponses et l'insertion des phrases de question.
- **Approche basée sur les templates :** Elle se réfère à l'utilisation de modèles pré-définis pour générer des questions à partir des dépendances entre les mots d'une phrase. Ces modèles sont conçus pour extraire des informations spécifiques des dépendances grammaticales, telles que les relations de sujet-objet, de modification, ou d'autres relations syntaxiques, afin de générer des questions appropriées.

4.4.2 La combinaison de plusieurs approches : Cette approche utilise les deux approches basées sur la sémantique et la syntaxe pour la génération de questions. récemment, **Dhole et Manning.**[12] ont développé Syn-QG, ils ont utilisé une collection de règles sémantiques, syntaxiques et heuristiques pour convertir une phrase déclarative à un couple question/réponse.

4.5 La génération des distracteurs : La génération de distracteurs peut être un processus complexe qui dépend du domaine et du type de données impliquées, mais implique généralement l'analyse des données d'entrée, la génération de mots similaires, la vérification de la pertinence et de la difficulté, et la sélection finale des distracteurs appropriés.

Il existe plusieurs approches pour sélectionner des mots clés pertinents à partir d'un texte qui comprennent :

4.5.1 Informations sur les parties du discours (part-of-speech information) :

Le distracteur est sémantiquement près de la clé. Par conséquent, la clé et les distracteurs doivent appartenir à la même catégorie de parties du discours. [1]

4.5.2 Word Embedding :

Le word embedding désigne un ensemble de techniques de machine learning qui visent à représenter les mots ou les phrases d'un texte par des vecteurs de nombres réels décrits dans un modèle vectoriel (ou Vector Space Model). L'idée derrière les word embeddings est que les mots qui ont des significations similaires soient situés à proximité les uns des autres dans cet espace.[13]

Il existe plusieurs algorithmes populaires pour apprendre les word embeddings, tels que:

- **Word2Vec**:

  Cet algorithme est parmi les plus connus. Il a été développé par une équipe de recherche de Google sous la direction de Tomas Mikolov. Il repose sur des réseaux de neurones à deux couches et cherche à apprendre les représentations vectorielles des mots composant un texte, de telle sorte que les mots qui partagent des contextes similaires soient représentés par des vecteurs numériques proches. Il possède deux architectures neuronales, appelées:

  - **Continuous Bag of Words(CBOW):** cette architecture reçoit en entrée le contexte d'un mot, c'est-à-dire les termes qui l'entourent dans une phrase, et essaye de prédire le mot en question.
  - **Skip-Gram** : cette architecture fait exactement le contraire de CBOW, elle prend en entrée un mot et essaie de prédire son contexte. Dans les deux cas, l'entraînement du réseau se fait en parcourant le texte fourni

et en modifiant les poids neuronaux afin de réduire l'erreur de prédiction de l'algorithme. [13]

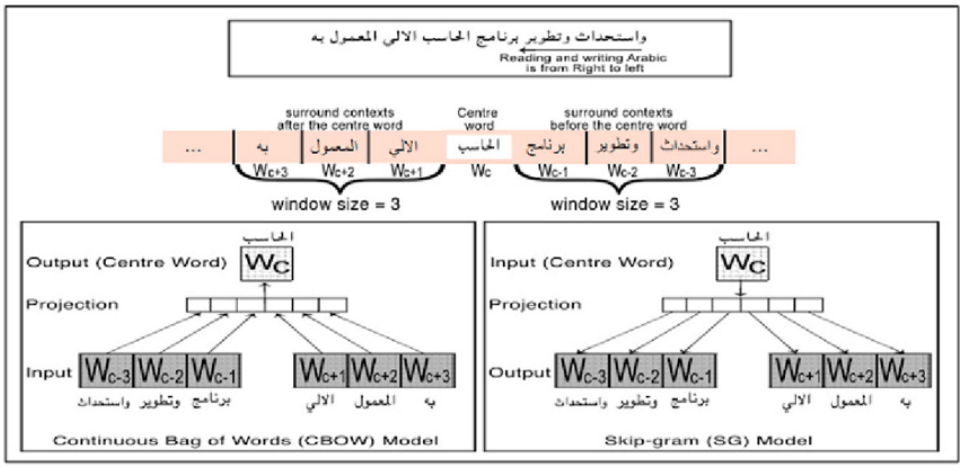

Figure 1.3 : Architecture de l'algorithme Word2Vec[14]

- **Glove (Global Vectors For Word Representation) :**
  Est un algorithme d'apprentissage automatique non supervisé utilisé pour générer des embeddings de mots, qui sont des représentations vectorielles de mots dans un espace de haute dimension. Bien que GloVe ait été initialement développé pour la langue anglaise, il peut également être utilisé pour d'autres langues, y comprisl'arabe, en entraînant le modèle sur un corpus de texte en arabe.

4.6 Post Traitement :

Le post-traitement est la phase finale qui vise à améliorer la qualité des QCM générés par le système. Différents types d'erreurs peuvent être présents dans les QCM générés par le système. Ceux-ci incluent erreur de ponctuation, mot interrogatif inapproprié, radical trop long, disponibilité des mots du discours en question, erreur de nombre accord, mauvaise qualité des distracteurs. Le système devrait minimiser ces erreurs à partir de la sortie finale. L'étape de post-traitement essaie de rectifier ces erreurs, sinon, supprime les erreurs des questions. Trois types de post-traitement ont été principalement utilisés dans la littérature. Ce sont:

4.6.1 La post-édition : est une étape où un éditeur humain révise et corrige les questions produites par un système de génération de question automatique.

Heilman et al.[11], ont appliqué quelques étapes de post-édition, telles que le changement des points finaux en points d'interrogation et la suppression des symboles d'espacement superflus.

4.6.2 Classement des questions : le classement de questions est une étape où les questions générées sont classées selon leur pertinence et leur qualité. Cette étape est

importante pour sélectionner les questions les plus pertinentes et les plus intéressantes pour l'utilisateur final.

4.6.3 Filtrage des questions inacceptables : Heilman et al. [11] ont inclus un filtre en tant qu'étape de post-traitement. Cette étape a filtré les questions contenant uniquement des groupes nominaux composés exclusivement de déterminants, les groupes nominaux contenant des pronoms non résolus, ainsi que les questions dont la longueur était supérieure à 30 jetons.

5. L'évaluation des systèmes de génération de questions

L'évaluation des systèmes de génération de questions est une tâche cruciale pour mesurer la performance et l'efficacité du système. En effet, il est important de s'assurer que les questions générées sont pertinentes, non ambiguës et qu'elles mesurent correctement les connaissances et compétences des apprenants. Il existe deux types d'évaluation pour les systèmes de génération de questions : l'évaluation automatique et l'évaluation humaine que nous allons présenter ci -dessous.

5.2 Evaluation automatique

L'évaluation automatique des modèles de génération de questions est une tâche difficile à cause du manques de mesure conçues spécifiquement, cette méthode utilise les métriques pour mesurer la qualité des questions comme:

5.2.1 La précision :

La précision est une mesure couramment utilisée pour évaluer la qualité des systèmes de génération de questions. Elle est définie comme la proportion de questions générées qui sont considérées comme correctes par rapport à un ensemble de questions de référence. En d'autres termes c'est le nombre de positifs bien prédit (Vrai Positif) divisé par l'ensemble des positifs prédit (Vrai Positif + Faux Positif). Cela nous donne sous forme mathématique :

Précision = Vrai positif / Vrai Positif + Faux Positif

où Précision = Qg ∩ Qh / Qg avec :

Qg : Les questions générées par le système

Qh : Les questions générées par l'être humain.

5.2.2 Recall :

La méthode d'évaluation Recall dans la génération de questions consiste à mesurer la capacité d'un système de génération de questions à générer des questions pertinentes à

partir d'un texte donné. Pour évaluer le Recall d'un système de génération de questions, on mesure la proportion de ces éléments qui ont été correctement identifiés et pour lesquels des questions ont été générées. En d'autres termes c'est le nombre de positifs bien prédit (Vrai Positif) par l'ensemble des positifs (Vrai Positif + Faux négatif). Sous forme mathématique :

Recall = Vrai positif / Vrai positif + Faux négatif

où Recall = Qg ∩ Qh / Qh avec :

Qg : Les questions générées par le système

Qh : Les questions générées par l'être humain.

5.2.3 F-mesure :

La méthode d'évaluation F-mesure dans la génération de questions est une mesure qui combine la précision et le Recall pour donner une mesure globale de la qualité d'un système de génération de questions. La F-mesure est calculée en utilisant la formule :

F-mesure = 2 * (précision * recall) / (précision + recall)

5.3 Evaluation manuelle

Dans cette section, nous allons présenter les différents critères permettant de mesurer la qualité des questions générées par les différentes approches de génération de questions manuelles.

a.La difficulté de la question : Il s'agit de donner une valeur aussi objective que possible pour définir la difficulté de la question.

b. Le pouvoir discriminant de la question : Il s'agit d'identifier sa valeur ajoutée par rapport aux autres questions générées.

c. L'utilité de chaque distracteur : Il s'agit de déterminer de façon globale dans quelle mesure chacune des mauvaises réponses proposées est pertinente vis-à-vis de la question.

d. La pertinence de la question: Ce critère consiste à vérifier si la question obtenue est spécifique à un domaine prédéterminé.

e. La qualité de la bonne réponse : Regroupant les notions de validité et de pertinence par rapport à l'énoncé.

f. La qualité grammaticale de l'énoncé : Ce critère est utilisé pour évaluer la qualité grammaticale de l'énoncé, donc les erreurs de langage et de syntaxe qu'il pourrait contenir.

g. La qualité informative de l'énoncé : ce critère est utilisé pour juger de la complétude de l'énoncé.

6. Les défis de la langue Arabe dans le domaine de TAL

La langue arabe fait face à de nombreux défis en raison de ses structures flexionnelles, dérivationnelles et syntaxiques. La grammaire de la phrase arabe a de nombreuses structures, principalement, la phrase nominale composée de sujet-verbe-objet (SVO), et la phrase verbale se compose de verbe-sujet-objet (VSO) . Outre la complexité de la langue arabe, les différentes représentations, l'interprétation sémantique et l'hétérogénéité des types de données sont souvent les principaux problèmes et les propriétés intrinsèques des données massives collectées,nous notons aussi ce qui suit :

- **Manque de ressources :** il y a moins de ressources disponibles pour le TAL arabe, telles que des ensembles de données étiquetés, des modèles pré-formés et des documents de recherche. Cela rend plus difficile le développement et la formation de modèles de haute qualité.
- **Complexité morphologique :** La langue arabe est une langue très morphologiquement complexe, ce qui signifie qu'elle a un grand nombre de formes différentes pour chaque mot, selon sa fonction dans la phrase. Cela rend difficile pour les modèles de génération de questions de déterminer le bon mot à utiliser dans la phrase.
- **Variations dialectales :** la langue arabe est une de nombreuses variations dialectales qui ont leurs propres règles grammaticales, vocabulaires et prononciations.[3]
- **Ambiguïté inhérente aux entités nommées :**

  La plupart des noms propres arabes sont indiscernables des formes qui sont les noms communs et les adjectifs qui pourraient créer une ambiguïté.

  Par exemple, le nom "الجزيرة "(Aljazeera) peut être reconnu comme un nom de l'organisation ou un nom correspondant à l'île. Néanmoins, les noms arabe dérivés d'adjectifs sont généralement ambigus, ce qui présente un défi crucial pour certaines applications de la NLP arabe telles que Reconnaissance des entités nommées en arabe.

  Par exemple, considérons le mot "أمل" (Amal), qui signifie « espoir », et peut être confondu avec le nom d'une personne.

  Dans les deux phrases suivantes, le mot "Amal" signifie deux différents sens :

  "الشباب هم أمل البلاد" ce qui signifie : la jeunesse est l'espoir de la pays.

  " أمل بنت جميلة " ce qui signifie : Amel est une belle fille.

7. Travaux connexes

Dans cette partie, nous allons présenter quelques travaux connexes que nous avons trouvés en ce qui concerne la génération automatique de questions à choix multiples en se basant sur les dépendances structurelles. Nous allons préciser l'approche utilisée pour la génération des questions et des distracteurs, les datasets utilisés, ainsi que les méthodes d'évaluation des questions générées.

➢ AQG: Arabic Question Generator [14]

Les auteurs de [5] ont commencé la conception de leur système par l'analyse d'un texte en arabe à l'aide de l'analyseur morphologique en ligne MADAMIRA qui effectue la segmentation en mots pas en phrase alors pour compenser cela ils ont intégré l'outil STAR (Arabic Text Slicer) en tant que module pour la segmentation de texte en phrases. Le résultat est un fichier segmenté et annoté en XML.

Ensuite AQG procède à l'attribution des différents rôles des acteurs de la phrase en utilisant des SRL propbank.

Expérimentation :

Dataset : 600 questions ont été dérivées de livres pour enfants et de résumés de livres, et ont toutes été générées par des humains.

Evaluation :

- Evaluation humaine : en considérant les critères (grammaticalité, réponse).

- Evaluation automatique : en calculant F-mesure. Les résultats étaient différents selon le texte sur lequel les questions ont été générées.

➢ Semantic Attributes Model for Automatic Generation of Multiple Choice Questions [15]

Les auteurs ont proposé un modèle de génération de question à choix multiple, l'algorithme suivant explique les bases suivies dans le modèle proposé:

- Construir la base de connaissances en extrayant les phrases qui ont les attributs sémantiques du jeu de données
- Sélectionnez une phrase interrogative et identifiez le type sémantique du mot clé en l'analysant sémantiquement
- Pour chaque phrase interrogative:
  - Mesurer la similarité entre la phrase interrogative et toutes les phrases de la base de connaissances.
- Triez les valeurs de similarité obtenues.
- Renvoie les trois phrases qui ont des valeurs de similarité plus élevées.

● Renvoie trois mots-clés des trois phrases comme distracteurs et identifier leurs types.

Expérimentation :

Dataset : TREC 2007 pour la réponse aux questions ,un ensemble de fichiers de différents sujets de domaine est analysé et 109 les phrases sont extraites pour être utilisées dans le test de la proposition maquette.

Evaluation: Pour évaluer la l'utilisation des algorithmes dans la génération des distracteurs, ils ont proposé quatre classes pour le niveau de difficulté de la question en utilisant les 8 algorithmes suivants: **N-gram, Smith, Levenshtein, Jaro, Cosine, Dice coefficient, Block Distance, Jaccard** .

- Une question est très difficile si tous les distracteurs générés ont le même type que le mot-clé.
- Une question est difficile si deux des distracteurs générés ont le même type que le mot-clé.
- Une question est de difficulté intermédiaire si une seule des distracteurs générés a le même type que le mot-clé. Une question est considérée comme facile si tous les distracteurs générés sont de types différents de celui du mot clé.

➢ **Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment [16]**

Dans ce système les auteur ont utilisé un pipeline pour générer des phrases simples ensuite, les phrases simples sont classées en fonction des topic-words pour sélectionner des informations sur les phrase informatives pour créer la racine du QCM, topic-words sont identifiés à l'aide de l'extraction automatique de mots clés (RAKE) . La technique de génération de distracteurs est proposée ici en utilisant le clustering non supervisé basé sur les fonctionnalités. Enfin, la similarité de chaîne et la similarité sémantique sont explorées au sein des clusters pour sélectionner les distracteurs finaux, qui sont les plus proches au corrigé.

Expérimentation :

Dataset : Il n'y a pas de dataset de référence disponibles pour évaluer ce système, Par conséquent, ils ont créé un ensemble de données de test pour vérifier ses performances. Le dataset a été créé en extrayant les pages Web de 14 dirigeants indiens et de 11 réseaux sociaux indiens. Le corpus de test comprend 25 documents composés de 1893 phrases.

Évaluation:

La précision du système proposé est évaluée à l'aide de l'équation suivante:

Précision (ACC) =(T P +TN) / TP + FP + FN + TN

où TP est le taux de vrais positifs, FP est le Taux de faux positif, FN est le taux de faux négatifs et TN est le taux de vrais négatifs

➢ **Automatic Distractor Generation for Multiple Choice Questions in Standard Tests [17]**

Les auteurs ont utilisé un réseau basé séquence à séquence pour générer les distracteurs. Ce cadre global contient cinq composants. Dans un premier temps, ils utilisent le codage module pour extraire les représentations sémantiques contextuelles pour tous les matériaux. Ensuite, ils utilisent l'attention mécanisme pour enrichir les représentations sémantiques de la question et de sa réponse. Enfin, concevoir trois composants clés pour générer des distracteurs utiles(question and answer guided Distractor Generation (EDGE) framework).

Expérimentation :

Dataset : RACE qui est collecté à partir des Examens d'anglais

(https://github.com/Evan-Gao/Distractor-Generation-RACE)

Évaluation:

- Pour l'évaluation de ce système les deux types d'évaluation ont été utilisées (automatique et manuelle).
- Pour l'évaluation manuelle, ils utilisent trois métriques conçues par Zhou et al. (2019) pour mener l'évaluation :
    - **la fluidité :** qui évalue si le distracteur suit la grammaire anglaise régulière et se conforme à la logique humaine et avoir le bon sens.
    - **La cohérence :** qui mesure si les phrases clés des distracteurs sont pertinentes pour le passage et la question.
    - **Capacité de distraction :** qui évalue la probabilité qu'un distracteur généré sera utilisé par les compositeurs de questions lors d'examens réels.

➢ **An Automated Multiple-Choice Question Generation Using Natural Language Processing Techniques [18]**

MCQG est un système utilisé pour traiter les supports de cours / documents alimentés par l'enseignant dans un choix multiple questions à côté des réponses à chaque question. Les processus NLP sont appliqués à l'aide de **TF-IDF et N-gramme**. Ce système ne s'applique qu'au texte du document ou du matériel de cours dans l'extraction mots-clés présentés par le professeur.

Expérimentation :

Dataset: cinq différents supports de cours de différentes tailles de phrases ont été sélectionnés.

Evaluation :

Pour l'évaluation des performances, les métriques de précision, de rappel et de f-mesure suivantes sont utilisées dans cette recherche.

Precision = numberCorrect /numberExtracted

Recall = numberCorrect/ totalExtracted

$\mathbf{F-measure} = 2 * (\mathit{Recall} * \mathit{Precision} / \mathit{Recall} + \mathit{Precision})$

8. Synthèse de travaux

Après avoir présenté les différents travaux connexes mentionnés précédemment, nous allons présenter ces derniers dans le tableau 1, en précisant l'approche utilisée, les métriques d'évaluation utilisées dans chaque travail ainsi que leurs langue et l'année de réalisation .

	Langue	**Année**	**Approche utilisée**	**métriques d'évaluation automatique**	**métriques d'évaluation manuelle**
[14]	Arabe	2020	Linguistique avec utilisation des rôles sémantiques	- Précision - Recall - F mesure	- Grammaticalité - Sens de la question
[15]	Anglais	2014	Étiquetage sémantique des rôles (SRL) et Reconnaissance d'entité nommée (NER)	- N-gram+ Smith - N-gram+ Jaccard	—

[16]	Anglais	2021	Tf_idf pour identifier les mots clés NER pour identifier la catégorie des mots clés K_means pour identifier les distracteurs	- Précision	—
[17]	Anglais	2020	séquence à séquence	- Rouge-L - Bleu4	- La fluidité - La cohérence - Capacité de distraction
[18]	Anglais	2021	Règles basées sur le balisage et les parties du discours.	- Précision - Recall - F mesure	- Difficulté - Pertinence - Exactitude - Ambiguïté

Tableau 1.1 : Résultats de comparaison des différents travaux.

9. Conclusion

L'évaluation est essentielle dans le processus d'E-learning et les QCM sont parmi les méthodes populaires pour l'évaluation pédagogique. Dans ce travail, nous avons passé en revue les travaux présentés dans le contexte de la génération automatique de QCM à partir d'un texte. Nous avons discuté des approches existantes pour la génération de QCM. Nous avons établi un générique de flux de travail à savoir le prétraitement, la sélection de phrases, la sélection de clé ,formation de questions, génération de distracteurs et les post-procédures. Nous présentons dans la phase suivante la méthodologie proposée pour construire notre système avant d'entamer la phase de mise en œuvre et la discussion des résultats.

Chapitre 2 : Conception

1. Introduction

Après avoir présenté l'état de l'art sur la génération de questions, nous présentons dans ce chapitre notre approche basée sur les dépendances structurelles pour la génération de questions à choix multiple en langue arabe. Donc d'après notre état des connaissances, nous allons explorer cette approche linguistique dans l'objectif d'améliorer la génération des questions. Dans la suite de notre travail, nous fournirons des détails sur l'architecture de notre système, avec les différents composants de ce dernier .

2. Définition de la tâche de génération de questions à choix multiple

La présente section a pour objectif de définir la tâche de génération de questions. Cette tâche revêt une grande importance dans le domaine du traitement de langage naturel.

Formellement, la tâche de génération de questions à choix multiple QCM consiste à produire une question Q = (q1, ..., qm) posée sur une réponse ciblée A, qui est une portion de texte dans le passage X = (x1, ..., xn). L'objectif est de proposer une question pertinente et intéressante pour l'apprenant tout en offrant plusieurs distracteurs qui permettent de différencier la réponse correcte des réponses potentielles fausses.

3. Approche proposée

Notre approche pour mener à bien notre travail repose sur l'utilisation de l'analyse syntaxique d'un texte qui consiste à identifier les relations grammaticales entre les mots du texte et des expressions régulières pour identifier des motifs spécifiques dans le texte. Ces deux techniques sont combinées pour obtenir des informations précieuses et extraire des éléments clés du texte.

Ensuite, nous utilisons ces relations et motifs identifiés pour générer des modèles de questions à choix multiple. Les mots clés pertinents sont sélectionnés en fonction des relations syntaxiques et des motifs spécifiques que nous avons définis à l'aide des

expressions régulières. Les modèles de questions servent de base pour formuler des questions qui captent les informations essentielles du texte.

De plus, nous générons également des distracteurs en utilisant les mots clés et les word embedding. Les distracteurs sont des options incorrectes mais plausibles qui peuvent tromper les apprenants lorsqu'ils répondent aux questions à choix multiple.

L'architecture de notre système est illustrée par la figure 2.1, où nous pouvons voir les différents composants qui le constituent. Chaque composant joue un rôle spécifique dans le processus de génération de questions à choix multiple, en utilisant les informations extraites du texte.

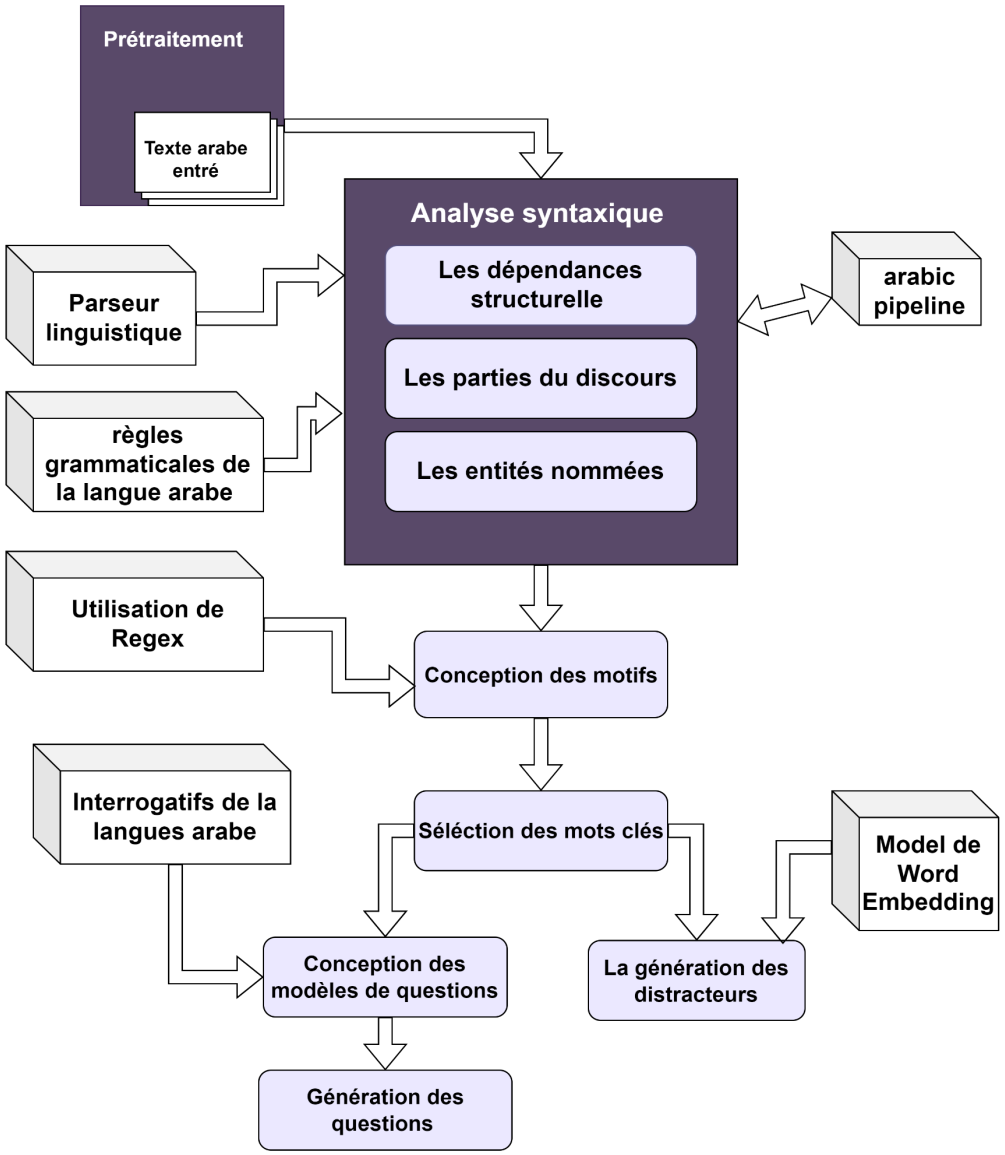

Figure 2.1 : Architecture de notre système de génération de questions.

Avant de commencer les processus de notre approche, il est essentiel de passer par une étape cruciale appelée prétraitement du texte. Le prétraitement consiste à préparer et à nettoyer le texte brut afin de le rendre plus adapté aux tâches de traitement du langage naturel. Cette étape revêt une grande importance car elle permet d'améliorer la qualité et la cohérence des données textuelles avec lesquelles nous travaillons.

3.1 Pré-Traitement de texte arabe entré

La langue arabe est connue par sa complexité et ses particularités en termes de grammaire et de syntaxe. Nous avons donc adopté une approche pour traiter le texte en entrée et le préparer pour la génération des questions selon les étapes mentionnés dans la figure 2.2 suivante :

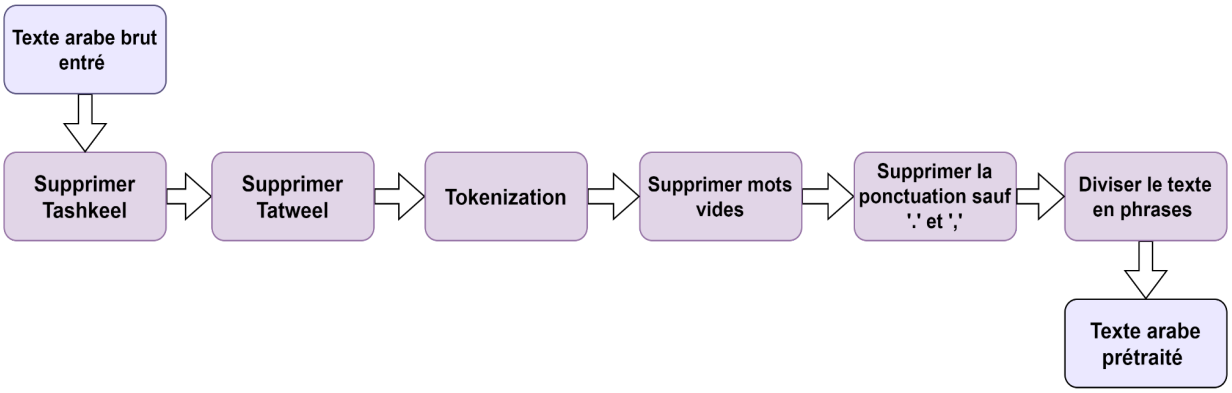

Figure 2.2 : Schéma illustrant les étapes suivies dans le prétraitement du texte.

Nous fournirons dans ce qui suit une explication détaillée des étapes mentionnées dans la figure 2:

a- Suppression des diacritiques (Tashkeel-التشكيل) :

La première étape consiste à supprimer les diacritiques du texte arabe entré. Les diacritiques sont des signes qui sont utilisés pour indiquer les voyelles de la langue arabe. Cette étape est importante car elle permet de normaliser le texte en le débarrassant de ces signes qui peuvent compliquer l'analyse et la compréhension.

Exemple :

Avant	Aprés
السَّماء	السماء

b- Suppression du "Tatweel" : éliminer le trait d'union arabe (tatweel) qui est utilisé pour prolonger l'écriture pour d'une lettre ou d'un mot

Exemple :

Avant	Aprés
العـــربية	العربية

c- Segmentation du texte (Tokenization)

Cette étape consiste à diviser le texte en mots individuels plus petits qui sont appelés tokens. Nous effectuons cette étape à l'aide de la bibliothèque "Pyarabic"[1].

أكل الاصدقاء التفاحة في البستان

['أكل', 'الاصدقاء', 'التفاحة', 'في', 'البستان']

Figure 2.3 : La Tokenisation d'une phrase à l'aide de " Pyarabic"

d- Filtrage des mots non pertinents :

La troisième étape consiste à filtrer les mots non pertinents du texte, qui peuvent mener à des difficultés d'analyse du texte.

أكل الاصدقاء التفاحة في البستان و ثم ذهبوا الى المدرسة

['أكل الاصدقاء التفاحة في البستان ذهبوا الى المدرسة']

Figure 2.4: Filtrage des mots non pertinents

f- Suppression des signes de ponctuation sauf le point et la virgule :

La cinquième étape consiste à supprimer tous les signes de ponctuation sauf la virgule et le point. Cette étape est importante car elle permet de simplifier le texte en supprimant les signes de ponctuation qui peuvent compliquer l'analyse et la compréhension.

[1] "Pyarabic" : Une bibliothèque spécifique pour la langue arabe en Python fournit des fonctions de base pour manipuler les lettres et le texte arabes.

يَستَيقِظُ زَيْدٌ مُبَكِّرًا!?!! يَتَوَضَّأُ فِي البَيْتِ ثُمَّ يَذْهَبُ إِلَى المَسْجِدِ

['يستيقظ زيد مبكرا ', ' يتوضأ في البيت يذهب إلى المسجد']

Figure 2.5: Suppression des signes "!" et "?"

g- Division du texte en phrases selon les signes de ponctuation :

La dernière étape consiste à diviser le texte en phrases selon les signes de ponctuation. Dans notre cas, nous avons choisi de diviser le texte en phrases selon les signes de ponctuation [".", "،", ","]. Cette étape est importante car elle permet de segmenter le texte en unités plus petites qui peuvent être analysées plus facilement.

ذهب الولد مسرعا. رجع الولد الى المنزل

↓

['ذهب الولد مسرعا ', ' رجع الولد الى المنزل']

Figure 2.6: La division d'un texte en phrases selon le signe de ponctuation "."

Une fois que nous avons exposé les différentes étapes du prétraitement de texte, nous commençons notre approche en effectuant une analyse syntaxique approfondie du texte en arabe qui nous est donné. Cette première étape revêt une grande importance car elle nous permet de comprendre la structure grammaticale du texte et les relations entre ses éléments. Nous allons ensuite expliquer plus en détail cette étape.

3.2 Analyse syntaxique d'un texte arabe

L'analyse syntaxique d'un texte arabe fait référence à l'étude de la structure grammaticale et des relations entre les mots dans une phrase ou un texte en langue arabe. Cela implique l'identification des différents éléments de la phrase, tels que les sujets, les verbes, les objets, les compléments, etc., et l'analyse de leurs fonctions syntaxiques. Elle permet de comprendre la manière dont les mots s'agencent dans une phrase et comment ils interagissent les uns avec les autres pour former un sens cohérent.

Dans cette phase nous utilisons la bibliothèque “Stanza”[2] pour l’analyse syntaxique. Nous initialisons le modèle en choisissant la langue arabe et les différents processeurs que nous souhaitons utiliser pour l’analyse du texte. Nous avons inclus les processeurs de l’analyse des dépendances structurelles et les parties de discours pour représenter les relations entre les mots de la phrase tels que le sujet, le verbe, l’objet, et les compléments, etc. De plus, nous avons également intégré les entités nommées pour identifier les noms des personnes et les locations présentes dans le texte.

Les processeurs choisis sont définis comme suit :

- **Les dépendances structurelles** : sont des relations grammaticales entre les mots d’une phrase qui décrivent la manière dont les mots se lient et interagissent les uns avec les autres. Nous introduisant dans le tableau 2.1 les différentes relations de dépendances que nous allons utiliser dans notre système

Relation de dépendance	Signification
NSUBJ	SUJET
OBJ	OBJET DIRECT
ROOT	Verbe ou nom فعل أوإسم/
OBL	complément circonstanciel de lieu / temps ظرف زمان/مكان
NMOD	modificateur nominal/المضاف إليه
AMOD	modifieur adjectives/ صفة أو حال
CCOMP	complément d’un verbe subordonné/ متمم لفعل متبع
XCOMP	complément d'un verbe non fini/ متمم لفعل غير مكتمل
ADVCL	complément adverbial /متمم الظرف

[2] “Stanza”: Stanza est une collection d'outils précis et efficaces pour l'analyse linguistique de nombreuses langues humaines. De la transformation du texte brut à l'analyse syntaxique et à la reconnaissance d'entités.

MARK	éléments de conjonction/ أداة ربط
CASE	éléments de préposition/حروف الجر
NUMMOD	représentation numérique/رقم أو عدد

Tableau 2.1: Relations de dépendances choisies dans notre système.

- **Les parties du discours :** Sont des catégories grammaticales qui classifient les mots en fonction de leur rôle syntaxique au sein d'une phrase. Elles permettent de décrire la fonction et la nature d'un mot dans un contexte linguistique donné. Le tableau 2.2 représente les parties du discours dans la langue arabe.

Les parties du discours	Signification
VERB	Verbe/ فعل
NOUN	Nom / إسم
NUM	Numéro / عدد
ADJ	Adjectif/ صفة
ADP	Préposition/ حرف جر
CCONJ / SCONJ	élément de conjonction / أدوات الربط

Tableau 2.2 : Représentation des parties de discours (Part-of-speech tagging) fournies par "stanza"

- **Les entités nommées:** Elles font référence à des éléments spécifiques et nommés tels que des personnes, des lieux et des organisations dans un texte donné. Elles représentent des informations importantes et identifiables qui peuvent être extraites pour une compréhension plus précise et une analyse plus approfondie du texte. Le tableau 2.3 fournit une description de ces entités nommées.

Entité nommée	Signification
LOC	Nom d'une place ou location/ اسم مكان
PER	Nom de personne / اسم شخص
ORG	Nom d'organisation / اسم هيئة أو منظمة

Tableau 2.3: Représentation des Entité nommée (Named Entity Recognition) fournies par "stanza".

Par la suite toutes ces représentation s'intègrent dans un pipeline pour fournir une séquence d'étapes qui sont appliqués sur le texte arabe, comme illustré dans la figure 2.7

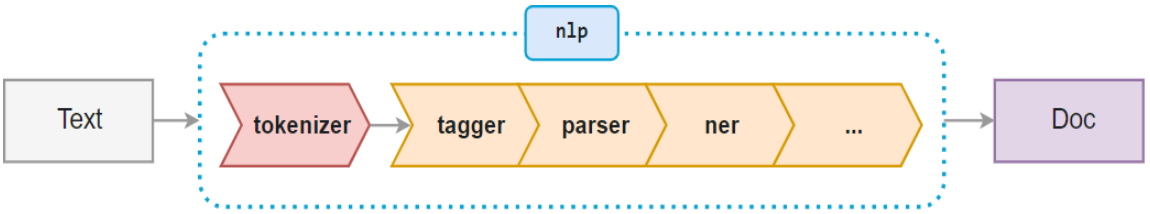

Figure 2.7: Représentation des caractéristiques linguistiques.[19]

Ensuite en utilisant ce pipeline nous allons obtenir une représentation analytique détaillée du texte arabe entré. Les figures **2.8 , 2.9 et 2.10** suivantes illustrent cette représentation .

La figure 2.8 représente le résultat fourni en effectuant les dépendances structurelles pour la phrase : ذهب أحمد إلى السوق

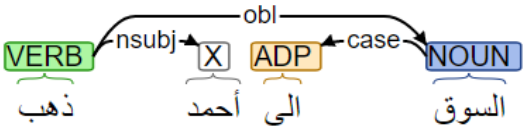

Figure 2.8: Exemple de représentation des dépendances structurelles fournie par "Stanza".[8]

La figure 2.9 représente le résultat fourni en effectuant les parties du discours pour la phrase : ذهب أحمد إلى السوق

Figure 2.9 : Exemple de représentation des parties du discours fournie par "Stanza".[8]

La figure 2.10 représente le résultat fourni en effectuant la reconnaissance des entités nommées pour la phrase : 1962 زار أحمد بلد الجزائر عام

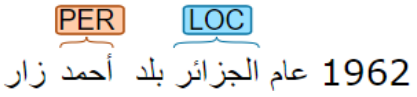

Figure 2.10: Exemple de représentation des entités nommées "Location" et "Personne" fournie par "Stanza".[8]

De plus nous introduisons les adverbes temporels (ظرف الزمان) et les adverbes de lieu (ظرف المكان) car ils sont tous les deux considérés comme un oblique (complément circonstanciel) dans les dépendances structurelles, sans faire la différence entre ces deux derniers. Nous essayons de les ranger dans une liste afin de pouvoir poser des questions de types temporels qui commencent par le mot (متى) et les questions sur les places qui commencent par le mot (أين).

Les deux figures 2.11 et 2.12 représentent les adverbes de temps et de lieu que nous utilisons dans notre système :

```
adv_temps_when = ['لدى' , 'عند' , 'بعد' , 'قبل' , 'الن'
          , 'جويلية' , 'جوان' , 'ماي' , 'أفريل' , 'مارس' , 'فيفري' , 'جانفي' , 'أيان' , 'إذا' , 'قط' , 'عندما' , 'بينما' , 'لدن'
          , 'الخميس' , 'الأربعاء' , 'الثلاثاء' , 'الأثنين' , 'الأحد' , 'السبت' , 'ديسمبر' , 'نوفمبر' , 'أكتوبر' , 'سبتمبر' , 'أوت'
          , 'شعبان' , 'جمادى الثانية' , 'جمادى الأولى' , 'ربيع الثاني' , 'ربيع الأول' , 'صفر' , 'محرم' , 'الجمعة'
          , 'مساءا' , 'ذو الحجة' , 'ذو القعدة' , 'شوال' , 'رمضان' ,
       'ظهرا' , 'فجرا' , 'متأخرا' , 'مبكرا' , 'باكرا' , 'ليال' , 'الليل' , 'صباحا' , 'المساء' , 'الصباح'
          , 'غدا' , 'البارحة' , 'أمس' , 'مغرب' , 'ربيع' , 'خريف' , 'شتاءا' , 'صيفا' , 'عصرا'
          , 'تموز' , 'حزيران' , 'أيار' , 'نيسان' , 'آذار' , 'شباط' , 'الثاني' ,
       'كانون الأول' , 'تشرين الثاني' , 'تشرين الأول' , 'أيلول' , 'آب' , 'الخريف' , 'الصيف' , 'الشتاء' , 'الربيع' ]
```

Figure 2.11: Représentation des adverbes temporels

```
place = ["خلف","جنوب","شمال", "غرب", "شرق"
        , "تحت", "ذراع", "ميل", "مكان", "ناحية", "جانب", "ذات" "فوق", "وراء", "قدام"
        , "حول" "شطر", "حيث", "ثم", "هنا", "أين", "يسار", "يمين",
        "بعد", "لدى", "لدن", "حوالي", "نحو", "أسفل", "أعلى", "بين" ]
```

Figure 2.12: Représentation des adverbes de lieu

Effectivement, afin de poursuivre la construction de notre modèle, nous avons développé des motifs d'expressions régulières spécifiquement pour la langue arabe. Dans ce qui suit, nous expliquons la façon de réaliser cette phase.
Mais avant de l'entamer nous devons d'abord définir les expressions régulières.

3.3 Les expressions régulières

Une expression régulière est une suite de caractères spéciaux que l'on nomme couramment « motif » ou « pattern » en anglais, et que l'on utilise dans de nombreux codes et langages informatiques pour lire, contrôler et assurer la validité du texte.
Les patterns de codes permettent de rechercher des chaînes de caractères ayant des propriétés communes dans un bloc de texte, et de leur appliquer un traitement automatique en les modifiant ou en les supprimant d'un seul coup, sans devoir faire les corrections une par une. Ce qui représente un gain de temps appréciable pour les développeurs de logiciels et d'applications.[20]
Les expressions régulières ou motifs Regex sont essentiellement un petit langage de programmation hautement spécialisé embarqué dans plusieurs langage de programmation tel que Python et dont la manipulation est rendue possible par l'utilisation du module re[3] [21]. En annexe, nous trouvons une description détaillée de regex.

Après avoir établi les définitions des expressions régulières et de leurs outils associés, nous procédons à présent à la conception des motifs. Cette étape vise à identifier les entités présentes dans le texte. Nous fournirons une explication détaillée par la suite.

[3] "re" : Le module re permet d'utiliser des expressions régulières avec Python.

3.4 La conceptions des motifs des expressions régulières

Ces motifs sont des expressions qui nous permettent d'identifier des schémas de caractères spécifiques dans le texte. Dans notre cas, ces motifs sont utilisés pour filtrer les entités pertinentes dans le texte, telles que les noms, les verbes, les adjectifs, et ainsi de suite.

Les motifs d'expressions régulières sont créés en se basant sur les règles linguistiques propres à la langue arabe, notamment la syntaxe. Ces règles nous permettent de définir les caractéristiques et les structures grammaticales des mots et des phrases dans la langue arabe. En comprenant ces règles, nous pouvons concevoir des motifs d'expressions régulières qui capturent les schémas linguistiques spécifiques de la langue arabe, ce qui nous permet d'extraire les entités pertinentes et d'enrichir notre modèle avec des informations linguistiques spécifiques à cette langue.

Dans le cadre de notre système, nous explorons les différents types de phrases et les différentes grammaires appliquées dans la langue arabe. En arabe, les types de phrases se divisent en deux catégories principales : les phrases verbales et les phrases nominales.

- **Définition d'une phrase verbale dans la langue arabe** : En langue arabe, une phrase verbale est appelée "جملة فعلية". Elle commence par un verbe فعل , qui peut être accompagné d'un sujet (فاعل) ou d'objets directs ou indirects (مفعول به).

  Exemple: أكل (فعل) الطفل (فاعل) تفاحة (مفعول به)

  Elle peut également inclure des compléments circonstanciels, tels que des adverbes ou des prépositions.

- **Définition d'une phrase nominale dans la langue arabe :**

  En langue arabe, une phrase nominale est appelée "جملة اسمية". Elle commence par un nom, elle est composée d'un sujet (المبتدأ) et un prédicat (الخبر)

  Exemple: الجو (مبتدأ) لطيف (خبر) , il existe plusieurs types de prédicat que l'on peut résumer comme suit :

  un prédicat unique: il se présentent sous la forme d'un mot explicite qui complète le sens

  Exemple: الأزهارُ متفتحةٌ

  Le prédicat est une phrase nominale : Ce prédicat comprend les composants de la phrase nominale principale, c'est-à-dire un sujet et un prédicat

  Exemple: الغرفة نظامها رائع.

Le prédicat est une phrase verbale : Ce prédicat se présente sous la forme d'une phrase qui contient les éléments de la phrase verbale traditionnelle, y compris le verbe, le sujet et son objet (le cas échéant).
Exemple: المؤمنة تجتهد في الخير
Le prédicat est la quasi-phrase : Ce type se présente sous la forme de la quasi-phrase du voisin et du neutre. Exemple : الفلاح في الحقل

Dans ce qui suit nous présentons les expressions régulières choisies dans notre système pour les deux types de phrase (verbale, nominale), chacune sera accompagnée d'un exemple et d'une figure qui représente l'analyse syntaxique de l'exemple donner . Nous commençons par présenter les phrases verbales, ensuite les phrases nominales.

3.4.1 Les expressions régulières pour les phrases verbales :

- Phrase verbale qui contient un verbe suivi d'un sujet :

  (<root><nsubj>)*

  Exemple : ذهب الولد

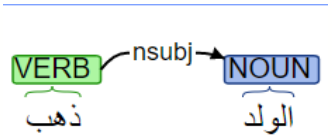

Figure 2.13 Exemple illustrant la relation de dépendance verbe-sujet.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet et un modificateur nominale :

  (<root><nsubj><nmod>)*

  Exemple : سافر صديق الولد

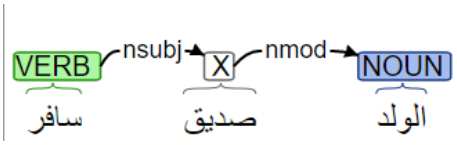

Figure 2.14: Exemple illustrant la relation de dépendances verbe-sujet-nmod.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet et un objet

  (<root><nsubj><obj>)*

  Exemple : يركب الولد السيارة.

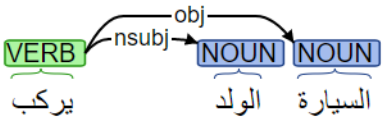

Figure 2.15: Exemple illustrant la relation de dépendances verbe-sujet-objet.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet, et un complément circonstanciel : **(<root><nsubj><case><obl>)***

  Exemple: ذهب الولد إلى الحديقة

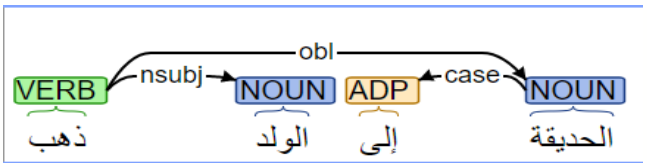

Figure 2.16: Exemple illustrant la relation de dépendances verbe-sujet-lieu.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet, un objet et un complément circonstanciel:

  (<root><nsubj><obj><case><obl>)*

  Exemple : قرأ التلميذ الكتاب في المكتبة

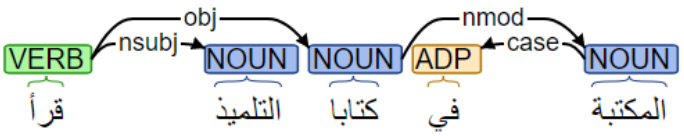

Figure 2.17: Exemple illustrant la relation de dépendances verbe-sujet-objet-lieu.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet, objet et modificateur nominal :

  (<root><nsubj><obj><nmod>)*

  Exemple: رأى الولد شجرة تفاح

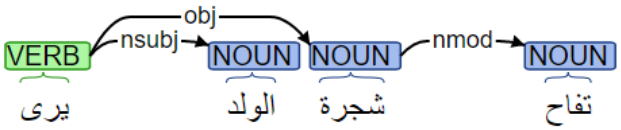

Figure 2.18: Exemple illustrant la relation de dépendances verbe-sujet-objet-modifieur nominal.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet, un objet, un modificateur nominal et un complément circonstanciel:

(<root><nsubj><obj><nmod><obl>)*

Exemple: رأى الولد شجرة تفاح في الحديقة

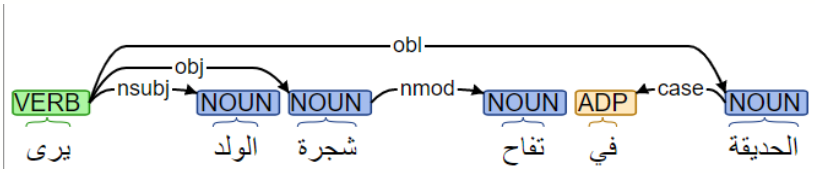

Figure 2.19: Exemple illustrant la relation de dépendances verbe-sujet-objet-modifieur_nominal_lieu.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet, un objet, un modificateur nominal, un adjectif et un complément circonstanciel :

(<root><nsubj><obj><nmod><amod><obl>)*

Exemple : رأى الولد شجرة تفاح كبيرة في الحديقة

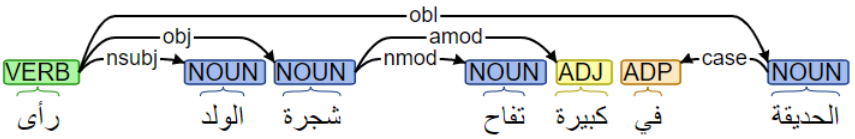

Figure 2.20: Exemple illustrant la relation de dépendances verbe_ sujet_ objet_ modifieur nominal_ adjectif lieu.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet,un complément circonstanciel (argument oblique) :

(<root><nsubj><case><obl:arg>)*

Exemple : ذهب الولد مع أخيه

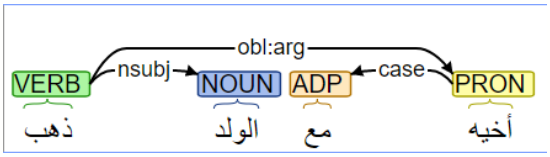

Figure 2.21: Exemple illustrant la relation de dépendances verbe-sujet_obl:arg.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet et un complément d'un verbe non fini :

(<root><nsubj><xcomp>)*

Exemple : بدأ الولد يقفز

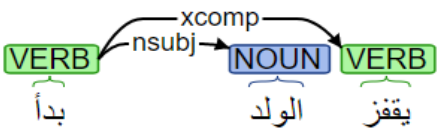

Figure 2.22: Exemple illustrant la relation de dépendance verbe-sujet-complément de verbe non fini.[8]

- Phrase verbale qui contient un verbe suivi d'un, un objet et complément d'un verbe non fini :

(<root><nsubj><obj><xcomp>)*

Exemple : شاهدت البنت القطة تمشي

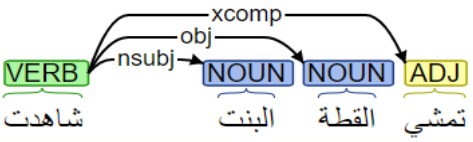

Figure 2.23: Exemple illustrant la relation de dépendances verbe-sujet-objet _complément de verbe non fini.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet, la conjonction de subordination "أن" suivi d'un complément d'un verbe subordonné:

(<root><nsubj><mark><ccomp>)*

Exemple : أراد الولد أن يذهب

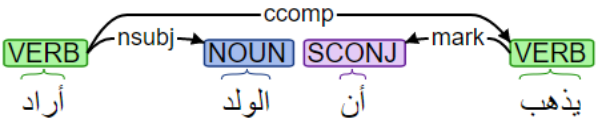

Figure 2.24: Exemple illustrant la relation de dépendances verbe-sujet_conjonction de subordination "أن" et complément d'un verbe subordonné.[8]

- Phrase verbale qui contient un verbe suivi d'un sujet , la conjonction de subordination "كي" qui exprime de subordination qui est utilisée pour introduire une proposition subordonnée dans le but, la finalité ou la conséquence , suivi d'un complément de verbe :

(<root><nsubj><mark><advcl>)*

Exemple : جاءت سلمى كي تساعدني

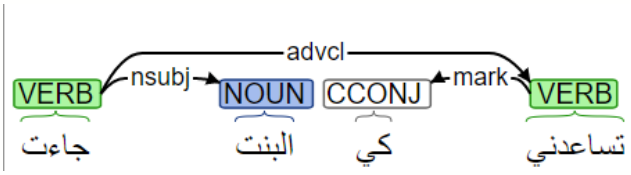

Figure 2.25: Exemple illustrant la relation de dépendances verbe-sujet-conjonction de subordination "كي" et un complément adverbiale.[8]

- Phrase verbale qui contient un verbe suivi un sujet, un complément circonstanciel et un chiffre ou un numéro :

(<root><nsubj><obl><nummod>)*

Exemple : 1962 استقلت الجزائر عام

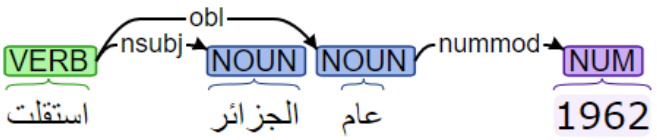

Figure 2.26: Exemple illustrant la relation de dépendances verbe,sujet,complément circonstanciel et un numéro.[8]

3.4.2 Les expressions régulières pour les phrases nominales :

- Phrase nominale qui contient un sujet et un prédicat (adjectif)

(<nsubj><root>)*

Exemple: الجو جميل

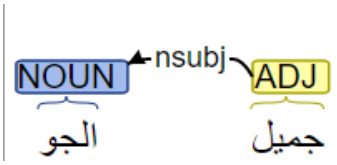

Figure 2.27: Exemple illustrant la relation de dépendances sujet et son prédicat.[8]

- Phrase nominale qui contient un sujet, un prédicat (adjectif) et un modificateur nominale:

(<root><amod><nmod>)*

Exemple: الجو جميل اليوم

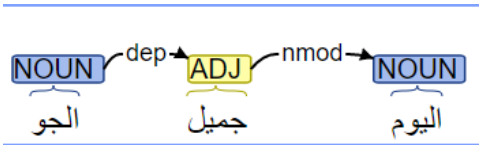

Figure 2.28: Exemple illustrant la relation de dépendances sujet et son prédicat.[8]

- Phrase nominale qui contient un sujet, un prédicat et un complément circonstanciel :

(<nsubj><root><obl>)*

Exemple : الجو جميل في المدرسة

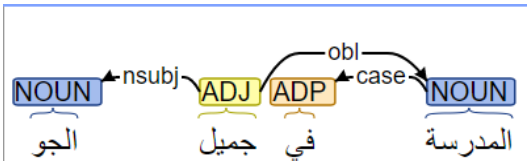

Figure 2.29: Exemple illustrant la relation de dépendances sujet et son prédicat et un complément circonstanciel.[8]

- Phrase nominale qui contient un sujet un verbe et un objet :

(<nsubj><root><obj>)*

Exemple : التلميذ يقرأ الكتاب

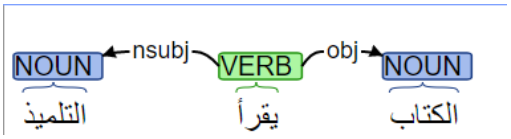

Figure 2.30: Exemple illustrant la relation de dépendances sujet-verbe-objet.[8]

- Phrase nominale qui contient un prédicat suivi d'un complément circonstanciel,un modificateur nominal et un adjectif :

(<root><case><nmod><amod>)*

Exemple : الطالب في المدرسة الجديدة

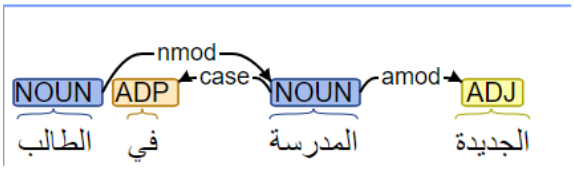

Figure 2.31: Exemple illustrant la relation de dépendances sujet-complément circonstanciel-adjectif.[8]

En effet, pour générer des questions à partir d'une phrase , il est nécessaire de passer par l'étape de la sélection des mots clés, Cette étape fait référence au processus de choisir un terme ou expression spécifique dans une phrase ou un texte. Dans ce qui suit nous montrons toutes les étapes suivies pour la sélection des mot clé.

3.5 La sélection du mot clé

Dans le processus de génération de questions, la sélection des mots clés revêt une importance primordiale. Elle consiste à identifier les éléments clés d'une phrase qui seront utilisés pour construire la question ainsi que la réponse. Dans notre système, nous utilisons les dépendances structurelles, les entités nommées et les parties du discours pour extraire ces mots clés pertinents. Pour ce faire, nous avons mis en place des expressions régulières spécifiques qui ciblent les relations de dépendance les plus courantes, telles que "root", "nsubj" et "obj" et les parties du discours tel que "verb", "noun", "adj", etc, ainsi que les entités nommées tel que "LOC" et "PER". Ces expressions régulières nous permettent d'extraire les sujets, les objets directs et les compléments circonstanciels de la phrase. Nous prenons également en compte d'autres dépendances, comme "nmod" et "case", afin d'inclure les prépositions et les modifieurs nominaux. En appliquant ces expressions régulières aux relations de syntaxiques de la phrase, nous sommes en mesure d'extraire les mots clés les plus pertinents pour générer des questions de haute qualité. Cette approche nous permet de capturer les éléments essentiels de la phrase et de les utiliser de manière précise pour construire des questions pertinentes. Dans ce qui suit nous fournirons un schéma et un exemple illustrons cette phase.

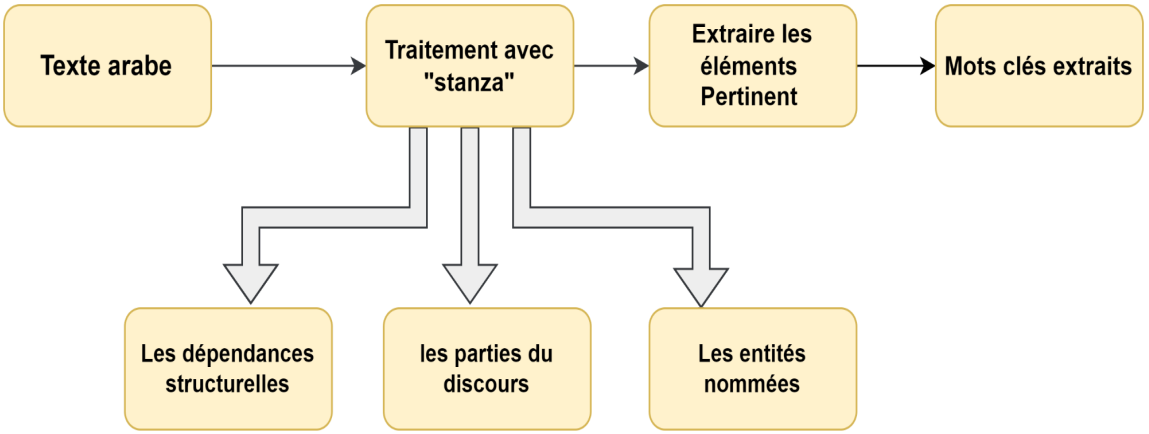

Figure 2.32 : Schéma illustrant la phase de sélection des mots clés

Exemple : "الولد يلعب في الحديقة"

En utilisant les dépendances structurelles, les expressions régulières ciblent la relation "nsubj" pour extraire "الولد" comme le sujet de la phrase qui est considéré comme un premier mot clé sélectionné dans le texte, ainsi que " verbe" pour extraire " يلعب" et la relation "obl" pour extraire "في الحديقة" comme un complément circonstanciel de lieu. Ces trois éléments pourraient être utilisés pour générer des questions correspondantes.

Après avoir extrait les mots clés et créé les motifs, nous entrons dans la phase de génération des modèles de questions. Pendant cette étape, nous prenons en compte à la fois les mots clés et les motifs afin de créer le modèle de question correspondant pour chaque mot clé. Nous veillons à ce que chaque question générée soit adaptée et appropriée au mot clé spécifique, en garantissant une correspondance pertinente entre le mot clé et la question formulée

3.6 La génération des modèles de questions

Cette étape consiste à formuler une question sur chaque mot clé généré en utilisant la structure de la phrase initiale et les relations de dépendance trouvées dans le texte ainsi que les expressions régulières. Pour cela, nous avons créé plusieurs modèles de questions, chacun adapté à une structure de phrase spécifique. Nous présentons dans le tableau 2.4 les modèles (templates) de questions proposés, chacun avec le motif approprié pour les phrases verbales.

Motifs appropriés pour les phrases verbales	Modèle de questions à générer
(<root><nsubj>)*	" من {verbe} ؟"
	" ماذا (ت)يفعل {sujet}؟"
(<root><nsubj><obj>)*	" من {verbe} ؟
	" ماذا (تفعل {sujet}؟"
	'ماذا {verbe} {sujet} ؟ '

(<root><nsubj><case><obl>)*	'من {verb} {preposition} {obl} ؟'
	'ماذا (ت)يفعل {subject} {preposition} {obl} ؟ '
	'این {verb} {subject} ؟ '
	'متی {verb} {subject} ؟ '
(<root><nsubj><obj><nmod>)*	'من {verb}{direct_obj}{nmod} ؟'
	'ماذا (ت)يفعل {subject} ؟ '
	'ماذا {verb} {subject} ؟ '
(<root><nsubj><obj><case><obl>)*	'من {verb} {direct_obj} {preposition} {location} ؟'
	'ماذا (ت)يفعل {subject} {preposition} {location} ؟ '
	'ماذا {verb} {subject} {preposition} {location} ؟ '
	'این {verb} {subject} {direct_obj} ؟'
	'متی {verb} {subject} {direct_obj} ؟'
(<root><nsubj><obj><nmod><case><obl>)*	'من {verb} {direct_obj} {nmod} {preposition} {obl} ؟'
	'ماذا {verb} {subject} {preposition} {obl} ؟'
	'ماذا (ت)يفعل {subject} {preposition} {obl} ؟ '
	'این {verb} {subject} {direct_obj} {nmod} ؟'
	'متی {verb} {subject} {direct_obj} {nmod} ؟'

(<root+<nsubj><obj><nmod><amod><case><obl>)*	'من {verb} {nmod} {preposition} {obl} {adj} ؟"
	'ماذا (ت)يفعل {subject} {preposition} {location} {adj} ؟ '
	' اين {verb} {subject} {direct_obj} {nmod} {adj} ؟"
	' متى {verb} {subject} {direct_obj} {nmod} {preposition} {obl} {adj} ؟"
	' كيف {verb} {subject} {direct_obj} {nmod} {preposition} {obl} ؟"
	' ماذا {verb} {subject} {preposition} {obl} ؟"
(<root><nsubj><xcomp>)*	‘؟ {subject} ماذا (ت)يفعل ‘
	‘ ؟ {xcomp} من الذي ‘
	’ يفعل (تفعل) ؟ {subject} {verb} ماذا‘
(<root><sujet><case><obl:arg>)*	' من الذي (التي) {verb} {preposition} {arg} ؟"
	'ماذا (ت)يفعل {subject} ؟ '
	' مع من {verb} {subject} ؟"
	‘بماذا {verb} {subject} ؟"
	'من ماذا {verb} {subject} ؟"
	' من اين {verb} {subject} ؟"
	' على ماذا {verb} {subject} ؟"

(<root><nsubj><obj><xcomp>)*	' ماذا (ت)يفعل {subject} ؟"
	' (التي) من الذي {verb} {direct_obj} {xcomp} ؟"
	' من الذي {xcomp} ؟"
	' ماذا {verb} {subject} ؟"
(<root><nsubj><mark><ccomp>)*	من' {verb} {mark} {ccomp} ؟"
	ماذا (ت)يفعل' {subject} ؟"
	ماذا" {verb} {subject} ؟"
(<root><nsubj><mark><advcl>)*	' ماذا (ت)يفعل {subject} ؟"
	' (التي) من الذي {verb} {markey} {advcl} ؟"
	' لماذا {verb} {subject} ؟"
(<root><nsubj><obl><nummod>)*	' {verb} سنة {nummod} من الذي أو التي؟"
	' ماذا وقع ل؟ {subject}'
	' متى {verb} {subject} ؟ '

Tableau 2.4 : Tableau résumant les motifs d'expressions régulières choisis dans les phrases verbales avec leur modèle de questions correspondant

Après la génération des modèles de questions, la sélection des mots clés, et à partir des expressions régulières et les dépendances structurelles, nous présentons le tableau 2.5 illustrant un exemple de questions générées pour un motif de questions et le mots clé généré pour la phrase verbale.

Texte	motif	mot clé généré	Question générés
رأى الرجل شجرة تفاح كبيرة في الحديقة	(<root><nsubj><obj><nmod><amod><case><obl>)*	رأى شجرة كبيرة	ماذا (ت)يفعل الرجل ؟
		الرجل	من رأى شجرة كبيرة ؟
		شجرة	ماذا رأى الرجل ؟
		كبيرة	كيف رأى الرجل شجرة تفاح في الحديقة ؟
		في الحديقة	أين رأى الرجل شجرة تفاح كبيرة ؟

Tableau 2.5 : Un exemple illustrant les questions et les mots clé générés pour le motif (<root><nsubj><obj><nmod><amod><case><obl>)*

Maintenant nous présentons les modèles de questions générés chacun avec le motif approprié pour les types de phrases nominales choisi précédemment dans le tableau 2.6:

Motifs appropriés pour les phrases nominales	Modèle de questions à générer
(<nsubj><root>)*	' (من الذي (التي) {prédicat}؟'
	' (هي) كيف هو {subject} ؟'
(<root><amod><nmod>)*	('تكون) متى يكون {subject} {adj} ؟'
	' اين {noun} {adj} ؟ '
	'ب (ت)وصف من {adj}؟'
	' (ت)وصف كيف {noun} ؟'
(<nsubj><root><obj>)*	'من {verb} {direct_obj} ؟'
	'ماذا {verb} {subject} ؟ '
	'يفعل(ت) ماذا {subject} ؟ '
	‘(هي) من هو {prédicat} ؟'

(<nsubj><root><obl>)*	‘(هي) كيف هو {subject} ؟'
	'اين {subject} {prédicat} ؟'
	'متى {subject} {prédicat} ؟'
	{subject} كيف وصف(ت) ؟
(<root><case><nmod><amod>)*	'من في {nmod} ؟'
	'اين {noun} ؟'

Tableau 2.6 : Tableau résumant les motifs d’expressions régulières choisis dans les phrases nominales ale avec leur modèle de questions correspondant

Après la génération des modèles de questions pour les phrases nominales, la sélection des mots clés à partir des expressions régulières et à partir des dépendances structurelles,

nous présentons dans le tableau 2.7 des exemple de questions générées pour l’un des modèles de questions :

Texte	motif	mot clé généré	Question générés
الجو جميل في المدرسة	(<nsubj><root><obl>)*	في المدرسة	اين الجو جميل ؟
		جميل	كيف هو (هي) الجو ؟
		جميل	كيف وصف(ت) الجو ؟

Tableau 2.7: Un exemple illustrant les questions générés pour le motif (<nsubj><root><obl>)*

Après la présentation des modèles de phrases nominales et verbales nous passons aux modèles de questions générés pour les entités nommés (Named Entity Recognition) qui sont illustrés dans le tableau 2.8 :

Texte	Entité	Question générés
الجزائر بلد سياحي جميل	Location : الجزائر	ما اسم المكان الذي ذكر في النص ؟
ذهب محمد إلى المدرسة	Person : محمد	ما اسم الشخص أو الأشخاص الذين ذكروا في النص ؟

Tableau 2.8 : Exemple de questions générés pour les entités nommées

Après avoir exposé les différents modèles de questions, nous entrons dans la phase cruciale de génération des distracteurs. Nous visons à créer des distracteurs convaincants qui semblent plausibles mais qui sont en réalité incorrects. Cela permettra de tester la compréhension et la perspicacité des apprenants, tout en favorisant une réflexion critique et une meilleure assimilation des connaissances.

3.7 La génération des distracteurs

Une tâche essentielle pour compléter la génération d'un questionnaire à choix multiple est la création de distracteurs. Les distracteurs sont des options de réponse erronées qui sont ajoutées aux questions à choix multiples pour augmenter la difficulté et évaluer la compréhension de l'utilisateur.

L'une des méthode utilisée pour créer ces distracteurs consiste à utiliser un modèle d'incorporation de mots pré-entraîné (word embedding). Les word embeddings sont des représentations vectorielles de mots qui capturent leur sens et leur contexte dans un espace vectoriel continu [22]. En utilisant un modèle de word embedding pré-entraîné pour une langue spécifique, il est possible de trouver des mots qui sont similaires en termes de sens et de contexte, et qui peuvent donc être utilisés comme distracteurs appropriés.

Il existe deux algorithmes d'apprentissage non supervisé pour les modèles de word embedding qui sont CBOW (Continuous Bag-of-Words) et Skip-gram.

- CBOW (Continuous Bag-of-Words) est un modèle qui vise à prédire un mot donné en fonction de son contexte (les mots environnants). Le modèle est nourri par le contexte, et prédit le mot cible.[22]
- Skip-gram est un model qui fonctionne dans le sens inverse de CBOW. Il est nourri par le mot cible, et prédit les mots du contexte. (les mots environnants).[22]

De plus, les word embedding sont caractérisés par la dimensionnalité. Elle fait référence au nombre de dimensions dans lesquelles la représentation vectorielle d'un mot est définie. Il s'agit généralement d'une valeur fixe déterminée lors de la création de l'incorporation de mots. La dimensionnalité de l'incorporation de mots représente le nombre total de caractéristiques qui sont encodées dans la représentation vectorielle.

Différentes méthodes de génération d'incorporations de mots peuvent donner lieu à une dimensionnalité différente. Le plus souvent, les incorporations de mots ont des dimensions comprises entre 50 et 300, bien que des dimensions plus élevées ou plus faibles soient également possibles. [23]

La figure 2.33 illustre le fonctionnement de modèle CBOW.

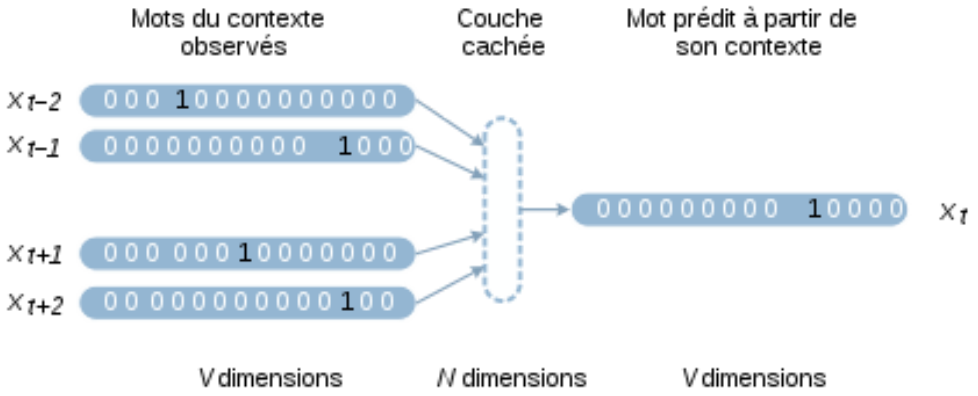

Figure 2.33 : L'architecture du modèle CBOW [24]

Et la figure 2.34 illustre le fonctionnement de modèle Skip Gram

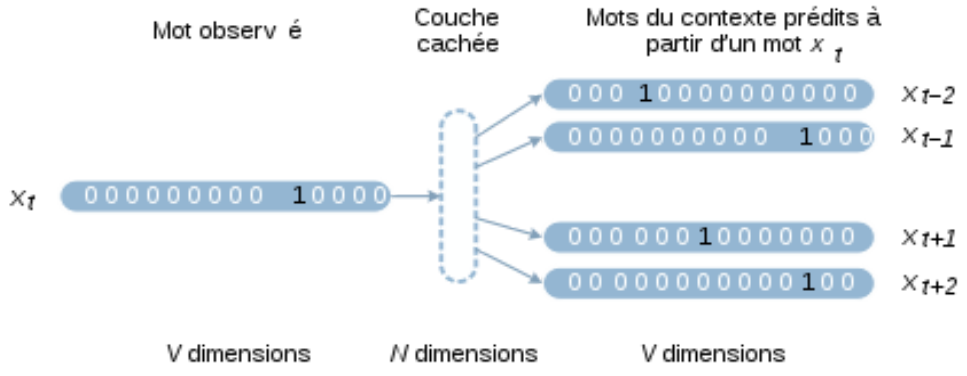

Figure 2.34 : L'architecture du modèle Skip Gram[24]

Pour générer des distracteurs à partir d'un modèle de word embedding pré-entraîné, plusieurs approches sont possibles. L'une d'entre elles consiste à trouver des mots qui sont similaires aux mots clés de la question en utilisant des mesures de similarité de cosinus ou d'autres mesures de distance vectorielle. Les mots qui ont une similarité élevée avec les mots clés de la question sont alors sélectionnés comme distracteurs potentiels.

Notre approche pour générer les distracteurs consiste à choisir le mot clé ou un terme d'une phrase donnée qui représente la réponse vraie de la question générée , ensuite nous devons trouver le mot le plus similaire en calculant la similarité cosinus entre le vecteur de ce mot-clé et les vecteurs de tous les autres mots du vocabulaire du modèle de word embeddings pre-entraînée. Enfin, les mots les plus similaires sont sélectionnés comme des distracteurs.

Prenons l'exemple suivant "الولد يصلي في المسجد" :

Dans notre approche nous sélectionnons le premier mot clé , " الولد" qui est considéré comme le sujet de la phrase (nsubj). En utilisant un modèle de word embedding ce terme est ensuite utilisé comme entrée pour générer des distracteurs.

Il convient de noter que l'utilisation de mesures de similarité, telles que la similarité cosinus, peut améliorer la précision de la similarité entre les textes.

- **La similarité de cosinus:** cette mesure courante calcule la similarité en utilisant l'angle entre les vecteurs de représentation de deux termes définie par l'équation:

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 2.35: Formule de cosinus [25]

Les résultats sont définis dans [-1,1], par contre la ressemblance entre deux textes, les valeurs compris dans [-1,0[seront rejeter, le 0 indiquera la non similarité et les

valeurs dans]0,1] seront considérés comme étant le taux de similarité entre A et B qui sont les deux documents en question.

Dans le tableau 2.9 ci-dessous nous présentons quelques exemples de distracteurs générés en utilisant l'approche précédente:

Question	mot clé	Distracteur générés
كيف صارت اللوحة ؟	رائعة	مبهرة ممتازة ورائعة جميلة
مع من رفض الصوص أن يرجع ؟	أمه	أبيها أمها جدته والدته
من الذي (التي) تدرس ؟	حنان	ورغدة نجوى هناء ماجدة

Tableau 2.9: Exemples de distracteurs générés

4. Conclusion

Ce chapitre a été consacré à l'architecture de notre système de génération de questions à choix multiples, en mettant l'accent sur l'utilisation des dépendances structurelles, des entités nommées et des parties du discours. Nous avons détaillé toutes les étapes et le travail effectué pour développer notre système.

Dans le prochain chapitre, nous aborderons l'implémentation et l'évaluation de notre système. Nous présenterons l'environnement de développement ainsi que toutes les ressources et les outils utilisés pour garantir l'efficacité et la puissance de notre système.

Chapitre 3 : Réalisation et Évaluation

1. Introduction

L'évaluation de notre système de génération de QCM est une étape importante pour garantir sa qualité et son efficacité dans les contextes éducatifs. A travers ce chapitre nous allons présenter les outils utilisés ainsi que l'environnement de développement de notre système. Par la suite nous fournirons les méthodes, les outils nécessaires et les critères d'évaluation utilisés pour mesurer la qualité des questions et des distracteurs générés. Nous allons également discuter des avantages et des limites de l'utilisation de notre système surtout dans le contexte de la langue arabe.

2. Environnement de développement et outils utilisés

Notre modèle est basé sur une approche linguistique qui ne nécessite pas de phase d'entraînement ou de test. Contrairement aux modèles qui se basent sur l'apprentissage automatique, notre approche utilise des règles linguistiques pour générer des questions et des distracteurs. Pour exécuter notre modèle nous avons choisi d'utiliser la plateforme en ligne Kaggle qui fournit un environnement de développement en ligne avec 30G de RAM et il fournit aussi des ressources puissantes sur le cloud et il permet d'utiliser un maximum de 30 heures de GPU par semaine.

- Le prétraitement de texte arabe entré est la première étape conçue dans notre modèle. Pour cela nous avons utilisé la bibliothèque Pyarabic. Cette bibliothèque de python est spécifique pour la langue arabe, elle fournit des fonctions de base pour manipuler le texte arabe, tout d'abord nous avons chargé la bibliothèque Pyarabic et importé les modules nécessaires. Ensuite nous avons créé une fonction qui prend en entrée un texte et applique plusieurs traitements sur ce dernier comme la suppression des diacritiques, la suppression du "Tatweel", la Segmentation du texte (Tokenization), Filtrage des mots non pertinents et enfin la suppression des signes de ponctuations sauf le point et la virgule.
- Dans la deuxième étape de notre approche linguistique, nous avons effectué une analyse syntaxique sur le texte prétraité dans la première étape en utilisant la bibliothèque Stanza. Cette bibliothèque fournit des outils pour l'analyse syntaxique

des textes dans différentes langues, y compris l'arabe. Nous avons utilisé les modèles de langue pré-entraînés pour le texte arabe fournis par “Stanza” afin d'obtenir des informations telles que les parties du discours, les entités nommées et les dépendances structurelles des mots dans notre texte.

-Dans le cadre de notre modèle, nous avons développé des motifs d'expressions régulières spécifiquement pour la langue arabe à l’aide de “regex”. L'outil "regex" fait référence à un ensemble de bibliothèques, modules ou fonctionnalités disponibles dans différents langages de programmation, qui permettent de travailler avec les expressions régulières.

Dans notre cas nous travaillons avec la bibliothèque “re” du langage de programmation python. Ces motifs sont des expressions qui nous permettent d'identifier des schémas de caractères particuliers dans un texte. Ils sont utilisés pour filtrer les entités pertinentes telles que les noms, les verbes, les adjectifs, etc.

La création de ces motifs d'expressions régulières repose sur les règles linguistiques propres à la langue arabe, notamment en ce qui concerne la syntaxe. Ces règles nous aident à définir les caractéristiques et les structures grammaticales des mots et des phrases dans la langue arabe.

Grâce à cette approche, nous sommes en mesure de concevoir des motifs d'expressions régulières qui ciblent les schémas linguistiques spécifiques de la langue arabe. Ces motifs nous permettent d'extraire les entités pertinentes et d'enrichir notre modèle avec des informations linguistiques spécifiques à cette langue.

- Pour que notre approche fonctionne, il est nécessaire que le texte suive les motifs d'expressions régulières que nous avons créés. En d'autres termes, le texte doit être compatible avec les motifs que nous avons définis pour identifier les entités pertinentes. Lorsque nous appliquons ces motifs sur le texte prétraité, à l'aide des dépendances structurelles, les parties du discours et la reconnaissance d’entités nommées nous pouvons extraire les mots clés pertinents tels que les noms, les verbes et les compléments circonstanciels, etc. Ces mots clés et ces motifs d’expressions régulières sont ensuite utilisés dans la conception des modèles de questions, chacun correspond à un mot clé spécifique avec l’utilisation des mots interrogatifs associés.

De plus, les noms identifiés peuvent comprendre des entités nommées telles que des personnes, des lieux ou des organisations, tandis que les verbes extraits peuvent indiquer des actions ou des processus importants dans le texte. Les compléments circonstanciels, quant à eux, apportent des informations supplémentaires sur le

contexte ou les circonstances entourant ces actions. Ainsi, il est essentiel que le texte se conforme aux motifs d'expressions régulières pour permettre une extraction précise des mots clés nécessaires à la génération des distracteurs..

- Nous avons utilisé la bibliothèque Python 'gensim' pour charger un modèle de Word2Vec, qui est une technique de vectorisation de mots permettant de représenter les mots sous forme de vecteurs dans un espace de dimension réduite. Cette approche nous a permis d'exploiter les fonctionnalités offertes par 'gensim' pour rechercher les mots similaires à un mot cible ou à une phrase cible.

Pour évaluer la proximité entre les mots du texte et ceux du vocabulaire, nous avons utilisé la mesure de similarité cosinus fournie par la bibliothèque 'sklearn'. Cela nous a permis d'identifier les mots les plus pertinents à utiliser comme distracteurs pour les questions. Dans le cadre de cette étape, nous avons testé deux modèles pré-entraînés spécifiques à la langue arabe.
Les modèles que nous avons testé sont "arwiki_20180420" de Wikipedia2Vec [26] et "ArabicConLL17" [27].
- "arwiki_20180420" est un modèle disponible en deux formats : fichier texte (.txt) et fichier binaire (.bin). Chaque fichier possède deux dimensions 100 et 300 Ce modèle a été créé en utilisant l'algorithme de Skip Gram.
- "ArabicConLL17" : est un modèle disponible en deux formats : fichier texte (.txt) et fichier binaire (.bin). Chaque fichier possède une dimension 100. Ce modèle a été créé en utilisant l'algorithme de Skip Gram.

Cependant, malgré l'utilisation de ces modèles pré-entraînés, nous avons constaté que les résultats obtenus n'étaient pas entièrement satisfaisants pour répondre à nos besoins spécifiques. Les limites de ces modèles pré-entraînés, notamment en ce qui concerne la sensibilité grammaticale et la pertinence sémantique dans le contexte de génération de distracteurs.
- En suivant toutes les étapes précédentes, nous obtenons notre système de génération de questions à choix multiple dans la langue arabe.
- Afin de faciliter l'utilisation de notre travail, nous avons développé une interface web en utilisant une interface de programmation applicative (API) appelée Starlette.

Starlette est une API basée sur le langage de programmation Python, qui nous permet de créer des services web de manière efficace et pratique.

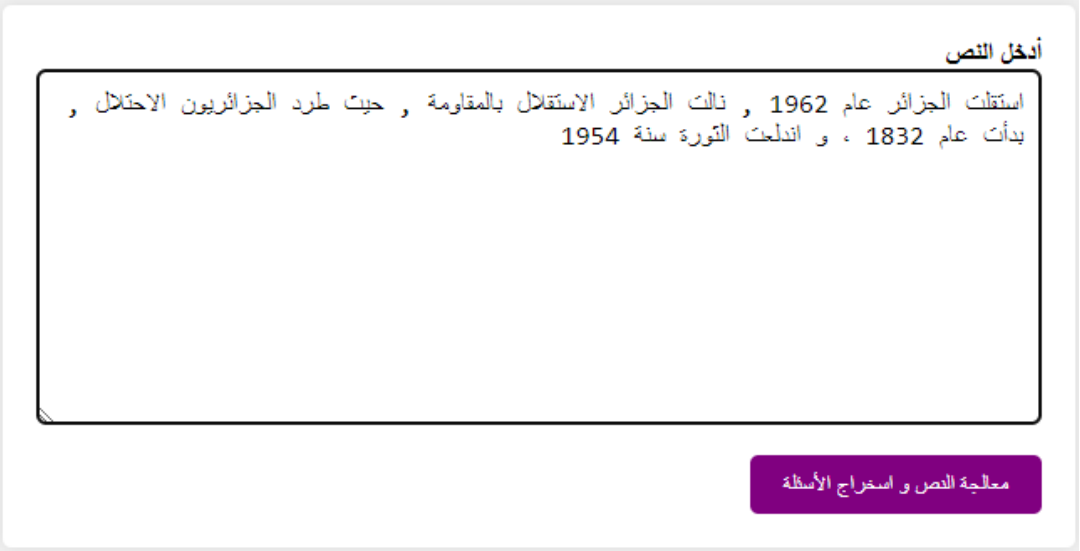

Figure 3.1: Interface de notre système

3. Evaluation de notre système

La plupart des études antérieures sur la génération de questions ont utilisé des métriques d'évaluation automatique pour mesurer les performances des modèles. Cependant, dans notre système, nous ne pouvons pas nous appuyer sur une évaluation automatique en raison de l'indisponibilité et de questions de référence adaptées à notre type de questions générées. Les questions générées par notre système appartiennent à un type spécifique qui n'est pas couvert par les questions de référence existantes en l'absence de dataset de référence puisque nous utilisons une approche non supervisée. Il peut s'agir, par exemple, de questions plus complexes ou de questions qui nécessitent un raisonnement plus avancé. Par conséquent, nous ne pouvons pas utiliser des métriques automatiques pour évaluer notre système.

L'idéal était de construire un jeu de test manuel pour l'utiliser dans cette évaluation. Cette tâche nécessite une expertise humaine particulière et fastidieuse que nous n'avons pas pu assurer.

Pour pallier cette limitation, nous avons dû adopter une approche d'évaluation manuelle.

3.1 Evaluation manuelle

Dans le but de procéder à une évaluation des performances de notre système, nous avons mis en place une évaluation humaine portant sur un échantillon[4] de cent (100) questions générées par notre système.

Pour mener cette évaluation nous avons sollicité l'expertise d'une personne compétente en langue arabe, qui a pris en considération deux aspects clés :

- Grammaire et Forme de question : Il est question de vérifier la structure des phrases, la concordance des temps verbaux, l'organisation des éléments de la phrase (leur place, leur ordre, éventuellement leur accord) et si les questions respectent les règles grammaticales de la langue arabe.

- Pertinence en sens de question : L'expert a évalué dans quelle mesure les questions générées étaient pertinentes par rapport au contexte donné. Il a examiné si les questions abordent les informations importantes ou si elles étaient hors sujet ou incomplète. Cette évaluation vise à déterminer si notre modèle était capable de générer des questions pertinentes en fonction du contenu fourni.

Les résultats d'évaluation des questions sont résumés dans le tableau 3.1 :

Modèle	Grammaire/ Forme de question /5	Pertinence en sens de la question /5
AMCQG	**4.00**	**3.50**

Tableau 3.1: Représentation des résultats de l'évaluation humaine des questions générées

En plus de l'évaluation des questions, l'évaluation des distracteurs est une étape cruciale pour assurer leur pertinence et leur efficacité dans le cadre de l'évaluation des connaissances. Dans notre processus, nous accordons une attention particulière à cette étape en procédant à une évaluation manuelle des distracteurs générés.

Pour cela, nous sélectionnons un échantillon représentatif de 100 distracteurs générés à partir de différentes questions. Ces distracteurs sont ensuite soumis à l'évaluation par l'expert de la langue arabe. Nous poursuivons l'évaluation avec les mêmes aspects de l'évaluation des questions, qui sont:

[4] https://huggingface.co/datasets/manelyasmine/Echantillon_devaluation/tree/main

- La grammaire / Forme des distracteurs : Cela signifie que les distracteurs doivent respecter les règles et la structure grammaticale de la langue arabe .

- Pertinence en sens de distracteur : Les distracteurs doivent être liés au sujet de la question, mais ils ne doivent pas être trop proches du mot clé pour éviter de rendre la réponse évidente.

Les résultats d'évaluation des distracteurs sont résumés dans le tableau 3.2 :

Modèle	Grammaire/ Forme de Distracteur/5	Pertinence en sens de Distracteur /5
AMCQG	**4.00**	**2.5**

Tableau 3.2: Représentation des résultats de l'évaluation humaine des distracteurs générés

Après avoir évalué les questions générées, nous avons constaté que l'un des principaux obstacles auxquels nous avons été confrontés est le manque de couverture exhaustive de tous les motifs de phrases existants en langue arabe et cela peut avoir un impact significatif sur la qualité et la diversité des questions générées.Cette limitation laisse encore un potentiel considérable pour de futures recherches visant à poursuivre et améliorer nos efforts. De plus, la génération de distracteurs a rencontré une pénurie importante en ce qui concerne les modèles de word embeddings. Ce qui limite l'efficacité de notre distracteur dans certaines situations.

4. Exemples de questions générées

Le tableau 3.3 suivant représente des exemples des questions à choix multiple générées par notre système.

Contexte	Questions générés	Mots clés	Distracteurs
تقع الجزائر في إفريقيا.تعتبر الجزائر بلدا سياحيا . الجزائر تتمتع بالمناظر الخلابة, والجبال الشاهقة, والشواطئ الجميلة. الجزائر غنية بالثروات الطبيعية	من الذي (التي) تقع في أفريقيا ؟	الجزائر	الجزائر/ENTITY قسنطينة وهران بالجزائر

تفع والواقعة هي وتقع	تقع	ماذا (ت)يفعل الجزائر ؟	
—	في افريقيا	أين تقع الجزائر ؟	
—	بلدا سياحيا	ماذا تعتبر الجزائر ؟	
—	بالمناظر	بماذا تتمتع الجزائر ؟	
والمرتفعات والتلال والجبال الوديان	الجبال	ماذا (ت)يفعل ؟	
والمرتفعات والتلال والجبال الوديان	الجبال	من وصف بالشاهقة ؟	
ورغدة نجوى هناء ماجدة	حنان	من الذي (التي) خرجت مع أمها ؟	خرجت حنان مع أمها ,و ذهبتا الى المجمع التجاري ,لشراء ثوب العيد .دخلت حنان دكان الملابس, فأعجبت بفستان زهري. دفعت الأم ثمن الفستان ,ثم عادت حنان فرحة الى البيت.
أختها أبيها جدتها والدتها	أمها	مع من خرجت حنان ؟	
—	الى المجمع	أين ذهبتا ؟	

دخلت دكان الألبسة دخلت دكان للملابس دخلت دكان الأحذية دخلت دكان والملابس	دخلت دكان الملابس	ماذا (ت)يفعل حنان ؟	
الإحتفال الأضحى عيد الاضحى	العيد	من ؟	
بيعها شرائها لشراء بيع	شراء	ماذا (ت)يفعل ثوب ؟	
ورغدة نجوى هناء ماجدة	حنان	من دخلت دكان الملابس؟	
سعدى ولبنى خولة ليلى	سلمى	من تدرس في كلية الفنون ؟	سلمى تدرس في كلية الفنون .قررت سلمى رسم لوحة,إستعملت الفنانة الأقلام في الرسم ,و صارت اللوحة رائعة
تدرس في كلية الأدائية تدرس في كلية التشكيلية تدرس في كلية للفنون تدرس في كلية والفنون	تدرس في كلية الفنون	ماذا (ت)يفعل سلمى ؟	
كلية الأدائية كلية التشكيلية كلية للفنون كلية والفنون	كلية الفنون	ما اسم المكان الذي ذكر في النص ؟	

قررت رسم لوحة قررت رسم اللوحة قررت رسم لوحات قررت رسم ولوحة	قررت رسم لوحة	ماذا (ت)يفعل سلمى ؟	
القديرة الراحلة المطربة والفنانة	الفنانة	من استعملت الأقلام الرسم في ؟	
فأصبحت وصارت وأصبحت أصبحت	صارت	ماذا (ت)يفعل اللوحة ؟	
مبهرة ممتازة ورائعة جميلة	رائعة	كيف صارت اللوحة ؟	

Tableau 3.3: Exemple de questions générés, mot clé et distracteurs

5. Conclusion

La langue arabe présente effectivement certains défis dans le domaine du traitement du langage naturel (TAL), notamment en ce qui concerne la pénurie des ressources linguistiques et des modèles pré-entraînés de word embedding.

Ces limitations ont posé des difficultés lors de la réalisation et de l'évaluation de notre système car l'outil que nous utilisons “stanza” présente certaines limitations en termes de qualité des résultats. Nous avons constaté que les résultats fournis par cet outil ne sont pas toujours optimaux et bénéficient de corrections supplémentaires. L'un des principaux problèmes que nous avons identifiés est la précision des analyses grammaticales et des reconnaissances d'entités nommées. Ces tâches peuvent être particulièrement complexes dans le cas de la langue arabe en raison de ses

particularités syntaxiques. Par conséquent, il arrive que les analyses grammaticales soient incorrectes ou que les entités nommées ne soient pas correctement traitées.

Ensuite en ce qui concerne l'évaluation de notre système nous n'avons pas pu effectuer l'évaluation automatique en raison de l'approche utilisée ainsi le manque des datasets de questions références. Cependant, malgré ces obstacles, nous avons pu obtenir des résultats acceptables et cohérents avec nos attentes initiales, en exploitant les ressources disponibles et en utilisant des techniques adaptées à la langue arabe.

Il est important de souligner que l'amélioration continue de l'état du langage naturel pour l'arabe est importante. Cela nécessite l'investissement dans la création de ressources linguistiques ainsi que le développement de modèles pré-entraînés.

Conclusion Générale

Notre travail vise principalement à automatiser la génération de questions à choix multiples en utilisant une approche linguistique basée sur l'analyse des dépendances structurelles (dependency parsing), l'étiquetage des parties du discours (POS tagging), la reconnaissance des entités nommées (NER) et la création de motifs de phrases à l'aide d'expressions régulières (Regex). Ces techniques nous permettent de générer automatiquement des questions à partir d'une phrase ou d'un texte en identifiant les concepts clés sur lesquels les transformations sont basées. Ensuite, nous avons abordé la génération des distracteurs en utilisant le calcul de similarités basées sur les cooccurrences incorporées dans les word embeddings issus d'un pré-entraînement profond.

Le principal avantage de notre approche par rapport à une approche supervisée est d'être plus facile à exploiter dans une variété de domaines puisqu'elle se base sur les constructions linguistiques et ne nécessite pas un ré-entraînement particulier pour prendre en considération les spécificités d'un nouveau domaine.

Pour évaluer l'efficacité et la performance de notre système, nous avons effectué une évaluation humaine en faisant appel à un expert en langue arabe. Nous avons pris en compte deux principales métriques d'évaluation : (la pertinence des questions et des distracteurs générées et leur grammaticalité). Il a été constaté une structure grammaticale très correcte et une pertinence qui bien acceptable doit être améliorée en faisant l'investissement de manière exhaustive de plus de variétés en termes de motifs linguistiques de la langue arabe.

Rappelons ici, la marge d'erreur imposée par le manque de pertinence des outils linguistiques utilisés.

Au cours de notre travail, nous avons rencontré certaines difficultés et problèmes:

- Nous avons remarqué que la langue arabe souffre d'une pénurie de ressources et d'outils par rapport à des langues plus courantes telles que l'anglais. Cette situation limite nos capacités à obtenir des résultats plus performants dans notre cas, notamment en ce qui concerne les outils d'analyse syntaxique de texte et la génération de distracteurs à l'aide de word embeddings.
- La langue arabe présente également une certaine ambiguïté, ce qui complique la résolution précise du sens des mots en fonction du contexte.

- La faible disponibilité de modèles pré-entraînés offrant des embeddings lexicaux riches a conduit à une importante lacune dans la génération de distracteurs pertinents. Cette partie doit être enrichie par une approche future plus aboutie.
- La non-disponibilité des questions de référence pour l'évaluation automatique nous contraint à effectuer une évaluation manuelle uniquement. L'expertise humaine étant un vrai challenge moyennant son indisponibilité.

Malgré ces difficultés, nous avons constaté que la communauté de recherche en domaine de traitement de langage poursuit ses efforts pour développer des méthodes et des ressources adaptées à la langue arabe. Cela nous a encouragés à surmonter ces obstacles et à choisir ce sujet de recherche prometteur.

Le travail que nous avons effectué jusqu'à présent représente une étape significative, qu' il est important de noter qu'il reste encore du travail en perspective dans deux dimensions:

- Enrichir les motifs pour couvrir le maximum des aspects grammaticaux de la langue ainsi que les modèles de génération de questions correspondants.
- Utiliser les modèles entraînés (BERT, T5, …) pour améliorer la qualité des distracteurs générés.

De plus, l'utilisation de techniques d'apprentissage automatique plus avancées, telles que l'apprentissage profond, pourrait aider à améliorer la performance de notre modèle dans la génération des questions à choix multiples.

Références bibliographiques

[1] D. R. CH et S. K. Saha, « Automatic Multiple Choice Question Generation From Text: A Survey », IEEE Transactions on Learning Technologies, vol. 13, no 1, p. 14–25, janv-mars 2020. https://doi.org/10.1109/tlt.2018.2889100.

[2] P. Sharma, « Dependency Parsing in Natural Language Processing with Examples » , *Analytics Vidhya*, déc. 2021, [En ligne]. Disponible sur : https://analyticsvidhya.com/blog/2021/12/dependency-parsing-in-natural-language-processing-with-examples/

[3] « I3rab Dependency Treebank » , *NLP Research Group Lab*. Consulté le 5 mars 2023. [En ligne]. Disponible sur : https://nlp.psut.edu.jo/malaac.html

[4] B. Zoulikha, « Université Ahmed Draia-Adrar : Intégration d'un lemmatiseur arabe dans Le cadre d'un système de recherche d'information », 21 mai 2019

[5] « The Stanford NLP Group » . https://nlp.stanford.edu/projects/arabic.shtml

[6] A. Pasha *et al.*, « MADAMIRA : A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic » , *Language Resources and Evaluation*, p. 1094‑1101, mai 2014, [En ligne]. Disponible sur : http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf

[7] K. Darwish et H. Mubarak, *Farasa : A New Fast and Accurate Arabic Word Segmenter*. Springer Science+Business Media, 2016, p. 1070‑1074. [En ligne]. Disponible sur : https://dblp.uni-trier.de/db/conf/lrec/lrec2016.html#DarwishM16/

[8] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, et C. D. Manning, *Stanza : A Python Natural Language Processing Toolkit for Many Human Languages*. 2020. doi : 10.18653/v1/2020.acl-demos.14.

[9] « Qu'est-ce que la reconnaissance d'entité nommée (ner) ? - définition de techopedia - l'audio 2023 » , *Icy Science*, 2023. https://fr.theastrologypage.com/named-entity-recognition

[10] N. Afzal et R. Mitkov, « Automatic generation of multiple choice questions using dependency-based semantic relations » , *Soft Computing*, vol. 18, no 7, p. 1269‑1281, juill. 2014, doi : 10.1007/s00500-013-1141-4.

[11] Michael Heilman, « Automatic Factual Question Generation from Text » CMU-LTI-11-004 Language Technologies Institute School of Computer Science Carnegie Mellon University 5000 Forbes Ave., Pittsburgh, PA 15213 www.lti.cs.cmu.edu

[12] K. Dhole et C. D. Manning, « Syn-QG : Syntactic and Shallow Semantic Rules for Question Generation » , *Meeting of the Association for Computational Linguistics*, avr. 2020, doi : 10.18653/v1/2020.acl-main.69.
[13] Data Analytics Post, « Word embedding » , *Data Analytics Post*, 13 décembre 2018. https://dataanalyticspost.com/Lexique/word-embedding/.
[14] K. Z. Bousmaha, N. H. Chergui, M. S. A. Mbarek, et L. H. Belguith, « AQG : Arabic Question Generator » , *Revue d'intelligence artificielle*, déc. 2020, doi : 10.18280/ria.340606.
[15] EldesokyFattoh, A. Elsayed Aboutabl, et M. Hassan Haggag, « Semantic Attributes Model for Automatic Generation of Multiple Choice Questions » , *International Journal of Computer Applications*, vol. 103, no 1, p. 18‑24, oct. 2014, doi : 10.5120/18038-8544.
[16] B. K. Das, M. Majumder, S. Phadikar, et A. A. Sekh, « Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment » , *Multimedia Tools and Applications*, vol. 80, no 21‑23, p. 31907‑31925, juill. 2021, Consulté le 27 février 2023. doi : 10.1007/s11042-021-11222-2.
[17] Z. Qiu, X. Wu, et W. Fan, « Automatic Distractor Generation for Multiple Choice Questions in Standard Tests » , *International Conference on Computational Linguistics*, déc. 2020, doi : 10.18653/v1/2020.coling-main.189.
[18] C. A. Nwafor et I. E. Onyenwe, « An Automated Multiple-Choice Question Generation using Natural Language Processing Techniques » , *International journal on natural language computing*, vol. 10, no 02, p. 1‑10, mars 2021, doi : 10.5121/ijnlc.2021.10201.
[19] A. Mipawa, « Understanding spacy process pipelines » , *NEUROTECH AFRICA*, juin 822, [En ligne]. https://blog.neurotech.africa/understand-spacy-pipeline/
[20] Data LabCenter, « Définition Expression régulière » , *Récupération de données au laboratoire Data LabCenter*, 2 décembre 2017.
https://www.data-labcenter.fr/glossaire/expression-reguliere/
[21] « Guide des expressions régulières » , *Python documentation*.
https://docs.python.org/fr/3/howto/regex.html
[22] « Word2vec : NLP & ; Word Embedding - DataScientest » , *Formation Data Science | DataScientest.com*, 16 mars 823. https://datascientest.com/nlp-word-embedding-word2vec
[23] B. O. C. Science et B. O. C. Science, « Dimensionality of Word Embeddings | Baeldung on Computer Science » , *Baeldung on Computer Science*, mars 2023, [En ligne]. Disponible sur :
https://www.baeldung.com/cs/dimensionality-word-embeddings#:~:text=Most%20commonly%2C%20word%20embeddings%20have,lower%20dimensions%20are%20also%20possible.

[24] Contributeurs aux projets Wikimedia, « Plongement lexical » , fr.wikipedia.org, mai 2023, [En ligne]. Disponible sur : https://fr.wikipedia.org/wiki/Plongement_lexical
[25] « Cosine Similarity Using Xilinx Alveo » , *Xilinx*. 7 mai 823 https://www.xilinx.com/developer/articles/cosine-similarity-using-xilinx-alveo.html
[26] Studio Ousia, « Pretrained Embeddings - Wikipedia2Vec » . https://wikipedia2vec.github.io/wikipedia2vec/pretrained/#arabic
[27] « NLPL word embeddings repository » . http://vectors.nlpl.eu/repository/#

Annexes

Dans cette partie nous introduisons la notion de “Regex”

- **Regex**

  Les expressions régulières (notées RE ou motifs regex) sont essentiellement un petit langage de programmation hautement spécialisé embarqué dans Python et dont la manipulation est rendue possible par l'utilisation du module re. En utilisant ce petit langage, nous définissons des règles pour spécifier une correspondance avec un ensemble souhaité de chaînes de caractères ; ces chaînes peuvent être des phrases, des adresses de courriel etc. [21]

  le tableau 4 ci dessous montre des exemple de caractères d’expressions régulières et leurs descriptions

  La manipulation des expressions régulières est possible sur Regex101.com. Les utilisateurs peuvent saisir un modèle d'expression régulière et le tester sur des chaînes de texte d'exemple pour voir s'il correspond ou extrait les informations souhaitées. Le site offre une représentation visuelle des correspondances, des explications sur les composants de la regex et des explications détaillées sur le processus de correspondance.

opérateur	description
.	Correspond à n'importe quel caractère unique de la ligne d'entrée.
^	représente le début de la ligne d'entrée.
?	Correspond à une chaîne de zéro ou un caractère qui correspondrait au caractère immédiatement à gauche de ?.
$	représente la fin d'une ligne d'entrée
\	Ce méta caractère est utilisé pour désactiver la signification spéciale des métacaractères.
[...]	Correspond à un ou plusieurs caractères ou

	à une plage de caractères de l'ensemble.
[^...]	Correspond à un ou plusieurs caractères ou à une plage de caractères ne faisant pas partie de l'ensemble
+	Correspond à une chaîne d'un ou plusieurs caractères qui correspondent au caractère immédiatement à gauche de +.
*	Correspond à une chaîne de zéro caractère ou plus qui correspondrait au caractère immédiatement à gauche de *.

Tableau 4 : Caractères d’expressions régulières avec leurs description