MINISTERE DE L'ENSEIGNEMENT SUPERIEURET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE SAAD DAHLEB – BLIDA 01

FACULTE DES SCIENCES

DEPARTEMENT D'INFORMATIQUE

MEMOIRE DE MASTER

Spécialité : Ingénierie des Logiciels

THEME

Vers un corpus monolingue parallèle de paraphrases pour la langue Arabe à partir de connaissances Web

Réalisé par:

- ✓ **Boukhatem Fatima Zahra**

Promotrice :

- ✓ **Mme OUAHRANI. L**

Membres de jury :

- ✓ **M. Hammouda** — **Président**
- ✓ **M. Riali** — **Examinateur**

Soutenu le : 03/10/2021

Remerciements

*J*e remercie tout d'abord DIEU le tout puissant qui m'a donné le courage, la patience pour faire ce travail.

*J*e tiens à remercier également toutes les personnes qui ont contribué de près ou de loin à l'élaboration de ce travail, et à la réalisation de ce mémoire.

*J*e tiens à remercier particulièrement Mme. Ouahrani Leila pour l'encadrement, le suivi du mémoire, ainsi que pour ses nombreux conseils, sa gentillesse et sa patience.

J'adresse mes remerciements aux membres du jury d'avoir fait l'honneur d'évaluer mon travail.

*U*n grand merci à mes chers parents pour leur amour, leurs conseils ainsi que leur soutien inconditionnel.

*J*e tiens à exprimer mes remerciements à tous les enseignants du département d'informatique à l'Université de Blida 1, spécialement M. Bala .M, pour leur patience et leur savoir faire.

J'adresse mes plus sincères remerciements à M. Khaled Boussebat qui m'a toujours soutenue et encouragé au cours de la réalisation de ce mémoire et tout au long de formation du master. Ainsi, Je tiens à remercier également M. Abid Dayaadine pour son aide et sa patience.

*Me*s remerciements vont également à M. Hadjaz Mohamed Bachir enseignant d'arabe à lycée boughani tizi ouzou & Mlle. Boukhatem Fawzia enseignante d'arabe à lycée abd el-hamid mahri ain defla pour leur précieuse collaboration en effectuant une expertise humaine sur nos jeux de données.

*J*e n'oublie pas mes chères sœurs (khawla, aicha, fawzia), mes chers frères (yassine, abd el madjid) et chers amis (lemya, chahinaz, imane) pour leur contribution, leur soutien et leurs encouragements.

ملخص

في الوقت الحاضر ، يتم استخدام اللغة العربية أكثر فأكثر على الويب. يمكننا أن نجد العديد من المقالات في جميع المجالات. ومع ذلك ، هناك القليل من الأعمال التي تستغل الويب كمصدر بيانات لأداء المهام المختلفة في البرمجة اللغوية العصبية.

يتطلب إنشاء تطبيقات جديدة في المعالجة التلقائية للغة العربية أولاً وقبل كل شيء تطوير نهج قوي وفعال.

في عملنا، نعالج مشكلة الإنشاء التلقائي لمجموعات أحادية اللغة من إعادة صياغة الجمل للغة العربية من اجل استخدامها لتدريب نموذج مخصص للغة العربية يقوم بصياغة الجمل تلقائيا.

يؤكد التقييم النوعي البشري على عينة من أزواج الجمل الناتجة عن نهجنا أن جودة إعادة الصياغة جيدة من الناحية اللغوية والنحوية.

يُظهر استخدام مجموعة البيانات الخاصة بنا لتدريب نموذج إعادة صياغة الجمل نتائج جيدة ، مما يسمح لنا باتخاذ خطوة كبيرة إلى الأمام في معالجة مشكلة ندرة الموارد العربية .

الكلمات المفتاحية: الويب ، البرمجة اللغوية العصبية ، إعادة الصياغة ، المجموعة أحادية اللغة ، مجموعة البيانات.

ABSTRACT

Nowadays, the Arabic language is used more and more on the web. We can find many articles in all areas. However, there is little work exploiting web knowledge as data resources to carry out the various tasks of NLP. The realization of new applications in automatic language processing (NLP) for Arabic requires first of all develops a powerful and robust approach.

In our work, we address the problem of automatic creation of monolingual corpora of paraphrases for the Arabic language in order to use them to train an automatic generation model of paraphrases dedicated to this language.

A human qualitative evaluation on a sample of pairs of sentences generated by our approach, confirms that the quality of the constructed paraphrases is good semantically and syntactically.

Using our dataset to train a paraphrase generation model shows tremendous results, allowing us to take a big step forward to address the Arab resource scarcity problem.

Keywords: web knowledge, NLP, paraphrases, monolingual corpus, Dataset.

Résumé

De nos jours, la langue arabe est de plus en plus utilisée sur le Web. On peut y trouver de nombreux articles dans tous les domaines. Cependant, il existe peu de travaux exploitant les connaissances Web comme ressources de données pour réaliser les différentes tâches du TAL. La réalisation de nouvelles applications en traitement automatique de la langue (TAL) pour l'arabe nécessite en premier lieu de développer une approche performante et robuste.

Dans notre travail, nous abordons le problème de création automatique de corpus monolingues de paraphrases pour la langue Arabe afin de les utiliser pour entrainer un modèle de génération automatique de paraphrases dédiée à cette langue.

Une évaluation qualitative humaine sur un échantillon de couples de phrases générées par notre approche, confirme que la qualité des paraphrases construites est bonne sémantiquement et syntaxiquement.

L'utilisation de notre dataset pour entrainer un modèle de génération de paraphrases montre des résultats considérables nous permet de faire un grand pas pour palier le problème de manque ressources arabe.

Mots clés : connaissances web, TAL, paraphrases, corpus monolingue**,** Dataset**.**

LISTE DES FIGURES

LISTE DES TABLEAUX

Liste des formules

Liste des acronymes

Acronyme	Déscription
TAL	Traitement Automatique de Langage
NLP	Natural Language Processing
Q&A	Question Answering
POS	Part Of Speech
WE	Word Embedding
EDAM	Encoder Decoder with Attention Mechanism
Seq2Seq	Modèle Séquence à Séquence
IR	Information Retrieval

Sommaire

INTRODUCTION GÉNÉRALE

1. Introduction

L'apparition de l'intelligence artificielle génère une source d'inspiration pour la majorité des chercheurs intéressés par le domaine de l'automatisation des techniques du langage pour modéliser un ensemble de règles syntagmatiques et linguistiques à base de l'ordinateur. Ces études imposent l'émergence d'une discipline nommée par le traitement automatique de la langue naturelle (TAL).

Les ressources linguistiques « corpus », représentant un outil indispensable pour toute sorte d'étude langagière car ils fournissent une base matérielle et un banc d'essai pour la construction des systèmes TAL. Cependant, une partie de l'attention a été concentrée sur les méthodes d'extraction des corpus ou encore les corpus de paraphrases.

Le concept de paraphrase est le plus généralement défini sur la base du principe d'équivalence sémantique i.e. des réalisations textuelles différentes de même sens. Dans la littérature, il existe deux approches de génération de paraphrases : l'une des approches est basée sur la transformation de textes (grammaires, transformations de surfaces de phrases, …) et l'autre consiste en des approches récentes guidées par les données (Data Driven).

De nombreuses approches basées sur les données ont été proposées. Les approches diffèrent considérablement par leur complexité et la quantité de ressources TAL sur lesquelles elles s'appuient. Parmi elles se trouvent des approches qui génèrent des paraphrases à partir d'un vaste corpus parallèle (contenant des couples de paraphrases) et reposent au minimum sur les outils de la TAL pour entraîner un modèle de génération ou d'identification de paragraphes. Les corpus parallèles sont utilisés dans plusieurs tâches notamment la formation (entraînement) de modèles séquences à séquences pour la génération et l'identification de paraphrases, la simplification de texte, la génération automatique de résumés, la traduction de textes, les systèmes de questions/réponses, la recherche d'information, …

2. Problématique

L'utilisation de corpus est indispensable dans le domaine de TAL et ses approches basées corpus.

Pour la langue arabe, l'utilisation de corpus de paraphrases entre dans le cadre de plusieurs contextes : évaluation automatique, calcul de similarité, recherche d'information…. Ceci fait apparaître quelques problèmes tels que :

- Le manque de corpus parallèles de paraphrases en langue arabe pour la génération et l'identification automatique de paraphrase. En effet, un manque de données et de recherches suffisantes affecte négativement les praticiens du traitement du langage naturel arabe.
- Le problème d'acquisition de corpus parallèles manuellement qui une tâche très difficile vu les tailles des corpus nécessaire pour les études de nos jours.
- Le problème de corpus parallèles de domaine général, peu de recherches ont concentrés sur l'acquisition de corpus spécifiques de domaine ou multi-domaine.

3. Objectifs

Notre objectif consiste à automatiser le processus de création d'un corpus monolingue parallèle de paraphrases pour la langue arabe. L'apport essentiel du travail proposé consiste à proposer une approche méthodologique pour la création d'un corpus monolingue parallèle de paraphrases, très peu exploité jusqu'ici pour l'arabe pour la génération automatique de paraphrases. L'approche proposée pour collecter automatiquement les données dans un corpus monolingue est une approche basée sur la collecte de connaissances Web.

Pour réaliser cet objectif notre travail consiste à :

- Explorer les approches de création de corpus pour les comparer et mettre en valeur notre approche,
- Elaborer une approche de conception automatique pour collecter et traiter les données.
- Mettre en œuvre l'approche proposée pour générer le corpus,
- Évaluer la qualité du corpus obtenu manuellement par expertise humaine,
- Évaluer le corpus par rapport à la tâche de génération automatique de paraphrases en l'utilisant pour entraîner un modèle de génération des paraphrases existant.

4. Importance du travail

L'importance de notre travail réalisé dans le cadre de ce projet de fin d'étude se résume dans les points suivants :

- Fournir un grand corpus parallèle monolingue de paraphrases pour la langue arabe pour entraîner un modèle de génération automatique de paraphrases.

- Automatiser le processus de création de corpus de paraphrases depuis le web.

5. Structure du mémoire

Les prochains chapitres sont structurés comme suit :

Le chapitre 1 présentera l'état de l'art du corpus monolingues de paraphrases ou nous allons introduire les différents concepts liés à notre travail. Dans le chapitre 2 nous allons présenter notre méthode de création de corpus de paraphrases, nous allons procéder à concevoir et à réaliser notre approche. Ainsi, nous allons effectuer une évaluation d'un modèle de génération de paraphrases en utilisant notre dataset.

Le dernier chapitre porte sur une consolidation d'outils qui permet d'automatiser le processus de création de corpus et nous terminons par une conclusion mettant le point sur le travail réalisé ainsi que des perspectives possibles.

Chapitre 1 : ÉTAT DE L'ART sur les corpus de paraphrases

1. Introduction

Décider si deux unités de texte ont la même signification est l'un des besoins les plus importants du traitement du langage naturel. Comme le langage naturel offre de nombreuses alternatives d'expression possibles, la capacité de déterminer que deux mots ou phrases ont une signification équivalente dans le contexte est nécessaire pour analyser le texte [1]. Dans la réponse à une question, par exemple, cela peut être utilisé pour extraire des réponses correctes exprimées avec des mots différents de ceux de la question. L'acquisition à grande échelle d'ensembles d'unités de texte équivalentes est un domaine de recherche actif.

Un certain nombre de techniques ont été proposées pour acquérir des unités de texte dans une relation de paraphrase, définie par une implication textuelle réciproque entre les deux unités.

Cette partie représente l'état de l'art sur les corpus de paraphrase dans laquelle nous allons présenter les différents concepts liés à notre contexte d'étude tels que la notion de paraphrase et la notion de corpus parallèle ainsi que les travaux connexes à notre travail. Nous allons structurer cette partie comme suit :

- Concepts fondamentaux ;
- Définition et classification des paraphrases ;
- Définition de corpus, corpus de paraphrase, corpus bien formé, les différents types de corpus et ces applications dans le domaine de TAL ;
- Les approches de construction de corpus parallèles avec une analyse comparative des méthodes existantes ;

2. Concepts fondamentaux

2.1. La langue ARABE et TAL

L'arabe est une langue chamito-sémitique. Elle fait partie de la branche sémitique avec l'hébreu et l'amharique en Éthiopie. Elle se présente sous deux formes principales : l'arabe *dialectal (vernaculaire)* et l'arabe *littéraire* (ou *classique*). L'arabe est largement parlé dans le monde aujourd'hui. La population arabophone est assez grande dans le monde avec ses

différents dialectes.

Le traitement automatique du langage naturel (TAL), ou NLP pour Naturel Langage Processing en anglais est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement de la langue naturelle (des corpus, des documents, des textes, des phrases, des mots, etc.) à la base des règles et grammaires linguistiques pour diverses applications.

D'une manière plus précise, le TAL traite de la conception de logiciels capables de traiter de façon automatique des données exprimées dans une langue (dite « naturelle », par opposition aux langages formels de la logique mathématique).

La langue arabe est parmi les langues sémitiques vivantes qui s'écrivent généralement de droite à gauche et qui est difficile à traiter par la machine. A la différence des autres langues comme, le français ou l'anglais, l'arabe est une langue très pauvre en termes de ressources et d'outils de traitement, bien qu'elle soit largement parlée dans le monde entier, mais ses ressources linguistiques utilisables dans le domaine du TAL sont rares à cause de certains facteurs :

1. La complexité de la langue (les propriétés morphologiques et syntaxiques), la présence de l'ambiguïté qui rend la langue -par fois- incompréhensible. D'après [2] « *l'ambiguïté est toute configuration linguistique dont la signification se construit par disjonction de deux termes mutuellement exclusifs*» C'est-à-dire le constituant linguistique en une seule forme correspondent plusieurs sens.
2. L'absence de voyelle a contribué à l'augmentation du taux de l'ambiguïté au niveau morphologique, syntaxique et sémantique comme par exemple le mot (شعر) : ce mot peut signifier : sentir شَعَرَ, poème شِعْرْ, cheveux شَعْرْ, etc.
3. Le vocabulaire riche de mots qui permet d'exprimer le même sens en utilisant des mots différents peut être aussi une source de difficulté, par exemple le mot « lion » peut avoir 500 autres mots qui l'expriment [3].
4. D'autres facteurs à ajouter sont l'absence des notions de majuscule et minuscule. Les lettres changent de forme en fonction de leur position dans le mot (début, milieu, fin, isolée). De plus, La structure d'un mot arabe est complexe, suite à l'agglutination de morphèmes lexicaux et grammaticaux. Un mot arabe peut désigner toute une phrase dans une autre langue ; par exemple le mot «أتتذكروننا» peut-être traduit en français avec « Vous souvenez-vous de nous ? ».

D’une manière générale, étant donné un mot ou une phrase, la difficulté de traitement automatique de la langue arabe réside dans le fait de savoir comment les outils de TAL peuvent déterminer le sens exact de ce mot ou cette phrase. C’est le problème qui motive les chercheurs en linguistique à développer une large gamme d'outils (TAL) pour analyser et annoter automatiquement différentes langues tels que les corpus, les lexiques, les dictionnaires, les jeux de données en plus des outils de fondamentaux entièrement automatisés tels que les tokeniseurs, les marqueurs de partie du discours, les analyseurs, les stemmers et les étiqueteurs de rôles sémantiques. ... [3]

L’étude de texte au niveau de phrases a également reçu de même l’intérêt des linguistes qu’ils ont fait apparaître la notion de « paraphrase » [4][5][6][7]. C’est l’étude qui permet de créer des ressources linguistiques pour le TAL en exploitant la relation de paraphrase entre deux ou plusieurs phrases pour déterminer d’une manière plus simplifiée leurs sens dans un contexte donné. *« La paraphrase peut aider à une facilitation de la compréhension du message par l’autre et contribuer au bon fonctionnement de la communication. »* [8]

2.2. La notion de paraphrase

Le mot **‘paraphrase’** dérive du latin **‘paraphrasis’** emprunté au grec **‘paraphrazein’**. Il est composé de **‘para’** (à côté de), et de **‘phrasis’ (**discours). Ce terme est apparu massivement dans la littérature linguistique depuis environ 15-20 ans.

Dans un sens général, paraphraser signifie reformuler. Il s’agit d’une opération qui consiste à changer les mots (la forme, la syntaxe) d’un texte afin de simplifier un texte, d’en faire un résumé, d’en faire une analyse sémantique ou reformuler des recherches web, etc. [9]

On dit que des séquences sont en relation paraphrastique (ou sont des paraphrases les unes des autres) quand elles ont « le même sens », « la même signification (la sémantique)», quand elles «veulent dire la même chose ».

Formellement, une paraphrase peut être modélisée comme une implication bidirectionnelle entre un texte **T** (texte d’origine) et une hypothèse **H** (la paraphrase) sous la forme, **T → H** et **H→ T** [10].

Exemple :

T : يقوم الهاكر باختراق الأنظمة من اجل الأموال

H : الدافع المادي للهاكر هو السبب الذي يجعله يخترق الأنظمة

Dans la pratique langagière quotidienne, le locuteur fait recours le plus souvent à la paraphrase pour plusieurs objectifs, lorsqu’il cherche des synonymes afin de varier son

discours, éviter les répétitions, reprendre les propos de l'autre ; en créant des énoncés sémantiquement équivalents.

La **figure 1**illustre la conception de Bernard Pottier [11] sur l'opération de produire des équivalences sémantiques, où :

Le niveau conceptuel représente l'aspect intentionnel où le lecteur veut dire quelque chose, puis il fait recours à son savoir (linguistique/encyclopédique) afin de choisir les mots qui conviennent avec ce qu'il veut dire (qui exprime son idée) en produisant un énoncé (niveau de langage naturelle) qui renvoie au (le dit 1), en relançant l'opération plusieurs fois, nous arrivons à un nombre infini d'énoncés de "redire"(le dit 2.....le dit n*): «on aboutit ainsi à un texte composé de plusieurs redites, toutes différentes entre elle, mais reliée à un même point de départ»*. [11]

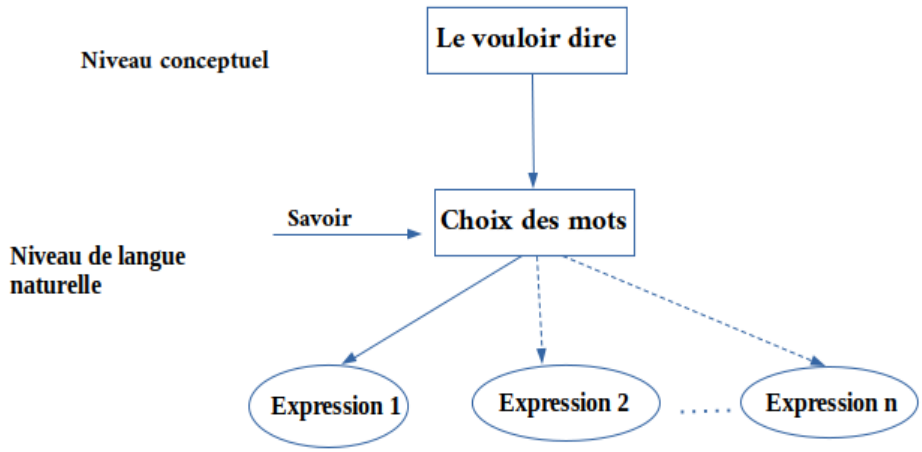


Figure 1 : l'opération de produire les équivalences sémantiques [11].

2.2.1. Classification des paraphrases

Les paraphrases peuvent être classifiées selon deux critères : le niveau de granularité et le niveau d'analyse de la langue.

2.2.1.1. Niveau de granularité

Le niveau de granularité des paraphrases concerne la taille des séquences, et on distingue deux types :

a. Paraphrase lexicale

Ce type concerne les séquences à un seul mot ayant un sens similaire. Ces séquences peuvent avoir une relation de synonymie tel que (paraphraser - reformuler) ou bien une relation d'hyperonymie qui présente la notion spécification/généralisation comme (humain, homme).

b. Paraphrase sous-phrastique

Les paraphrases sous phrastiques sont des fragments de texte (groupes de mots) présentant la même signification comme :

1 : زينب لا تكذب أبدا.

2 : زينب تقول الحقيقة دائما.

c. Paraphrase phrastique

Les paraphrases phrastiques sont des phrases qui transmettent le même sens en changeant seulement quelques mots et/ou passages.

Exemple :

1 : المخترقون هم من ينشرون الفيروسات في الأجهزة.

2 :. الفيروسات تزرع في الأجهزة من طرف المخترقين

2.2.1.2. Niveau d'analyse de la langue

La distinction des différents types de paraphrases au niveau d'analyse de la langue est basée sur la compréhension de la phrase et la détermination des conditions nécessaires pour cette tâche.

a. Paraphrase sémantique ou linguistique

La paraphrase linguistique se base sur les correspondances syntaxiques et ou lexicales entre les phrases. On distingue deux types à savoir : les paraphrases syntaxiques et les paraphrases lexico-syntaxiques.

i. **Paraphrase syntaxique**

Ce type de paraphrase se base sur une règle qui spécifie les conditions du passage d'une phrase à l'autre.

Exemple :

1 : لقد كان هذا السؤال صعبا لكنه استطاع الإجابة عنه

2 :استطاع الإجابة عن السؤال رغم انه كان صعبا لقد

Donc, plusieurs transformations peuvent être utilisées (la nominalisation, conversion d'un adjectif en syntagme nominal…) afin d'obtenir une paraphrase syntaxique.

ii. **Paraphrase lexico-syntaxique**

Dans ce type de paraphrases, les modifications peuvent être effectuées sur le niveau syntaxique et sur le niveau lexical.

Exemple :

1 : لقد فشل في الامتحان

2 : لم ينجح في الامتحان

b. Paraphrase non-linguistique

Les paraphrases non linguistiques se basent sur des phrases comportant la même idée ou référant la même chose sans chercher des correspondances lexicales ou syntaxiques ce qui nécessite l'intervention de l'expert.

b.1. Paraphrase pragmatique

Deux paraphrases pragmatiques sont des phrases qui réfèrent à la même intention de telle sorte que les phrases sont interprétées de la même façon en se basant sur l'expérience et les connaissances. Pour ce type de paraphrase l'équivalence sémantique est tributaire de la situation de communication.

Exemple :

1 : « يجب إن ارتدي معطفي »

2 : «الجو بارد اليوم »

Le deuxième énoncé n'est équivalent à 1 que dans une situation de communication déterminée avec des énonciateurs précis.

b.2. Paraphrase référentielle

Dans ce type de paraphrases, il est nécessaire de connaître les références de certains termes. Il faut une référence à la situation de nomination ce qui rappellera une valeur anaphorique, déictique...etc.

Exemple :

1 : النساء, الرجال والأطفال يريدون الذهاب

2 : الكل يريد الذهاب

الكل faits référencent à النساء, الرجال, الأطفال.

2.3. La détection de paraphrase

La détection de paraphrases est une tâche de TALN qui vise à identifier automatiquement si deux phrases ont la même signification. [12] définissent des assentiments paraphrases ou des phrases qui transmettent la même signification en utilisant des termes différents. De plus, ces

phrases représentent des formes de surface alternatives dans le même langage, exprimant le même contenu sémantique que les formes originales.

Plusieurs approches sont proposées pour la détection automatique de la paraphrase. Elles reposent sur les propriétés paradigmatiques des mots et sur leur capacité de se substituer mutuellement dans un contexte donné. L'une des méthodes la plus utilisée est la mesure de similarité syntaxique et sémantique entre deux chaines comme la méthode de mesure les chaînes d'édition dans un corpus monolingue qui permet de détecter les unités linguistiques (mots, syntagmes, etc.) permettent de rapprocher deux chaînes en basant sur les traits communs entre les deux unités [13].

Les corpus bilingues parallèles, qui contiennent la traduction d'un texte dans une autre langue, peuvent aussi être utilisés pour la détection de la paraphrase.

2.4. Corpus de paraphrases

2.4.1. Définition de corpus

Il existe de nombreuses définitions d'un corpus dans la littérature, prenons une définition souvent citée, celle de [14] : « *Un corpus est une collection de morceaux de langues qui sont sélectionnés et ordonnés selon des critères linguistiques et extralinguistiques explicites afin d'être utilisés comme un échantillon de la langue.* »

L'expression « morceaux de langues » est utilisée parce qu'un corpus peut contenir des textes ou des transcriptions des discours complets ou incomplets.

Selon un autre point de vue, plusieurs chercheurs ont considéré le corpus sous l'angle de la méthodologie de constitution [15], [16].

Les corpus sont des outils indispensables et précieux en TAL. Ils sont des collections de données sélectionnées et organisées selon des critères explicites pour servir d'échantillon pour un traitement particulier ou de référence permettant en effet d'extraire un ensemble d'informations utiles pour des traitements statistiques. Les corpus riches ont été aussi utilisés dans une variété d'applications telles que la segmentation, la discrétisation, étiquetage morphosyntaxique, désambiguïsation morphologique, segmentation des syntagmes et étiquetage des rôles thématiques (sémantiques).

Il existe plusieurs catégories de corpus disponible à l'utilisation, peuvent être général ou spécialisé et catégorisé comme suit :

- **Corpus de texte :** représentant les collections de documents textuelles brut tel quelles multilingues et monolingues, articles web et collections de chaque dialectale (Généralement des tweets).
- **Corpus annotés :** regroupés les corpus contenants des simplifications et des analyses regroupées sous forme d'ensembles sémantique, Part Of Speech(POS)[1], indication d'erreurs …. etc.
- **Lexique :** contenant des listes de mots et bases de données lexicales.
- **Corpus Manuel :** écrits à la main ou scannés et transformés en textes utilisables dans une machine.
- **Corpus de discours oral :** des textes retranscrit depuis des audio enregistrés
- **Divers :** cette catégorie regroupe les corpus divers du genre Questions/Answers (Q&A) ou la détection de plagiat pour certains cas d'utilisation.

2.4.2. Les types des corpus

• **Spécifique :** Un corpus contenant une collection de documents dont la relation est proche d'un domaine décrit par un ensemble de mot clés par l'utilisateur au moment de la génération.

• **Multi-Domaine :** Un corpus contenant plusieurs corpus spécifiques dont les documents sont tirés à partir d'une liste de mots clés également mais cette avec une relation entre eux de n-gram[2], pour garantir l'intégrité du corpus en termes de diversité de documents.

• **Générique :** Ce type de corpus représente une collection de documents qui n'appartient pas forcément au même domaine précis, les corpus de ce type peuvent contenir tous les documents que le système croisera durant son processus de génération.

2.4.3. Corpus bien formé

Pour qu'un corpus soit bien formé, plusieurs caractéristiques doivent être prises en considération lors de la création :

- **La taille :** le corpus doit atteindre une taille critique pour permettre des traitements statistiques fiables. Il est impossible d'extraire des informations fiables à partir d'un petit corpus.

[1]Part Of speech : chaque mot à une valeur propre à lui dans une phrase (nom, verbe, adj. etc.)
[2]Un ***n*-gram** est une sous-séquence de *n* éléments construite à partir d'une séquence donnée. L'idée était que, à partir d'une séquence de lettres données il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante.

- **Le langage du corpus :** un corpus bien formé doit couvrir un seul langage, et une seule déclinaison de ce langage.
- **Le temps couvert par les textes du corpus :** un corpus ne doit pas contenir de textes rédigés à des intervalles de temps trop larges.

2.4.4. Corpus de paraphrases

Les corpus de paraphrases sont des collections de paraphrases, qui consistent en des expressions linguistiques avec un libellé différent et (approximativement) la même signification [17]. On peut distinguer deux types de corpus :

Corpus parallèle : un corpus de paraphrases parallèle est un ensemble de textes (source) dont chacun est traduit dans une ou plusieurs langues autres (cible) que l'original. Le cas le plus simple est celui où seules deux langues sont impliquées : l'un des corpus est une traduction exacte de l'autre. Selon [18], un corpus parallèle est un corpus qui « se compose de textes originaux en langue source A et leurs versions traduites en langue B », tandis que [19] le décrit comme une « collection de textes, dont chacun est traduit dans une ou plusieurs autres langues ».

La **figure 2** représente un schéma d'un corpus parallèle où le corpus cible est le résultat de traduction d'un corpus source.

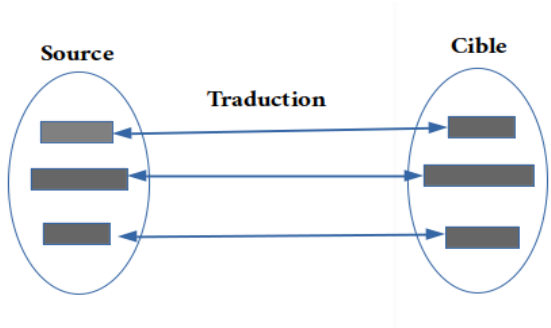


Figure 2 : Schéma d'un corpus parallèle [1].

Un corpus parallèle peut être multilingue où les paires de phrases sont disponibles dans deux langues ou plus [20] (comme les transcriptions des débats de parlement européens Europarl [21]). Et il peut être aussi monolingues où paires d'énoncés de sens équivalents sont alignées de façon supervisée (comme les groupes de questions ayant la même réponse) [22]. Les corpus parallèles unilingues sont utilisés pour créer et évaluer des systèmes guidés par les données pour le paraphrasage, la génération de texte et la simplification de texte.

Corpus comparable : les corpus comparables consistent en des ensembles de textes dans différentes langues qui ne sont pas des traductions les uns des autres. Le mot « comparable »

indique que le texte dans différentes langues a été sélectionné parce qu'il a certaines caractéristiques en commun. Par exemple, on sélectionne souvent des textes qui ont en commun : le sujet, la période ou le degré de technicité, etc.

Corpus monolingues comparables : des paires de textes associés en fonction d'une mesure de similarité textuelle suivant éventuellement certaines heuristiques.

Notamment, certains travaux exploitent des corpus monolingues comparables, comme ceux de [23] dans le domaine médical visant la construction d'un corpus de paraphrases de segments opposant les langues de spécialité et de vulgarisation.

2.4.5. Les différentes applications de corpus de paraphrases

Le corpus de paraphrases est un élément important dans le traitement de langage, il peut être utilisé pour entraîner plusieurs domaines tels que :

- **L'entraînement des machines de traduction :** [24] utilisent des paraphrases induites automatiquement pour améliorer un système de traduction automatique basé sur des phrases statistiques. Un tel système fonctionne en divisant un texte donné en phrases et en traduisant chaque phrase individuellement en recherchant sa traduction dans un tableau. La couverture du système de traduction est améliorée en permettant à toute phrase source qui n'a pas de traduction dans le tableau d'utiliser la traduction de l'une de ses paraphrases.
- **La génération automatique des résumés** dans le but d'améliorer les systèmes de résumés.
- **L'entraînement des systèmes questions/réponses :** c'est un domaine qui attire de plus en plus l'attention des chercheurs. Plusieurs travaux ont porté sur la construction de corpus de paraphrases de questions. [22] ont construit un corpus de paraphrases de questions en exploitant des réseaux sociaux sur le Web. Dans cette expérience, les auteurs ont recueilli un corpus de questions et leurs paraphrases à partir du site WikiAnswers[3].
- **La génération automatique des paraphrases :** les systèmes de génération ou identification automatique des paraphrases nécessitent un grand corpus de données afin de s'entraîner à produire des paraphrases pertinentes, comme dans [25] ou ils ont utilisé les données monolingues parallèles comme données d'apprentissage pour un modèle de traduction automatique Seq2Seq[4]pour former un traducteur à générer des paraphrases l'arabe à l'arabe.

[3]http://wiki.answers.com/

[4]Seq2Seq est une méthode de traduction automatique basée sur un codeur-décodeur qui mappe une entrée de séquence à une sortie de séquence avec une étiquette et une valeur d'attention.

3. Les approches de création de corpus parallèle

Ces dernières années, plusieurs corpus de données ont été développés pour être utilisés dans les applications du TAL et autres, en se basant sur plusieurs approches, toutes partagent le même but de fournir une grande quantité de données sous forme de corpus de paires de phrases. [4] ont distingué trois manières différentes pour la collecte de paraphrases. La première est l'utilisation de ressources linguistiques existantes. La seconde est l'extraction de mots ou d'expressions similaires en se basant sur un corpus. La troisième, enfin, est l'acquisition manuelle de paraphrases. Elle est sans doute la plus facile à implémenter, et ses résultats sont les plus fiables mais elle consomme beaucoup de temps.

a. Extraction de données à partir d'un corpus comparable

Cette approche a reçu beaucoup d'attention de la part de la communauté de chercheurs dans les années récentes. En effet, plusieurs techniques ont été proposées pour exploiter cette ressource, citant par exemple, les techniques de recherche qui utilisent des caractéristiques générales des documents (telles que le titre du document ; le lien du document, la structure du document, etc.) ainsi que des informations lexicales du document [26]. Un Corpus composé de textes dans la même langue partageant une partie du vocabulaire employé, ce qui implique généralement que les textes parlent d'un même sujet, durant la même période, peuvent contenir des paraphrases parallèles.

La construction des corpus de paraphrases parallèles à partir d'un corpus comparable consiste en appliquant une traduction multiple de ce dernier et en utilisant une langue pivot pour déterminer si les phrases obtenues sont réellement en relation de paraphrase.

Dans [27], les auteurs ont proposé une méthode non supervisée pour construire un corpus monolingue anglais à partir d'un corpus comparable. Quatre types de métriques de similarité de phrases ont été proposés, basés sur l'alignement entre word embeddings. Ils ont aligné chaque mot de la phrase complexe avec le mot le plus similaire de la phrase simple (une phrase simple est une reformulation d'une autre phrase dit complexe). Le corpus obtenu est composé de 492,493 phrases et 25 mots dans chaque phrase.

DSim [28] est un corpus parallèle bilingue danois aligné sur des phrases, extrait de 3701 paires de télégrammes d'information et de courts articles de presse simplifiés professionnellement correspondants. Le corpus est destiné à la simplification automatique du texte pour les lecteurs adultes. L'alignement des phrases se fait par le calcul de score de

similitude cosinus pondérée tf * idf[5] [29]. Les résultats d'alignement sont comparés à l'état de l'art pour l'alignement des phrases en anglais.

b. Extraction de données à partir d'un corpus bilingue

Cette approche consiste à traduire les phrases d'un corpus parallèle bilingue pour construire un corpus parallèle monolingue. Un système de traduction de la langue source vers la langue cible est utilisé pour traduire un corpus de la langue source, et former de nouvelles paires de phrases parallèles synthétisées source–cible. L'utilisation des corpus parallèles bilingues basé sur l'hypothèse d'équivalence de traduction, qui stipule que les unités de texte partageant des traductions dans au moins une autre langue peuvent être des paraphrases [20].

Le projet PPDB [30] a des ressources de paraphrases pour plusieurs langues, y compris l'arabe. Les paraphrases sont obtenues à l'aide de corpus bilingues parallèles en appliquant la méthode du pivot où une langue est utilisée comme pont ou représentation de signification intermédiaire.

L'article [31] décrit les efforts de LDC[6]dans la collecte, la création et le traitement de différents types de données linguistiques, y compris le texte parallèle. Les chercheurs ont utilisé la traduction de l'arabe et du chinois vers l'anglais pour construire un corpus parallèle.
Dans l'article [32], les auteurs ont présenté une méthode qui consiste à construire un corpus parallèle de paraphrases d'énoncés proposées par des contributeurs volontaires sous la forme de traductions multiples à partir de plusieurs langues européennes vers le français. Ils ont décrit quelques mesures dans le but d'évaluer le degré de similarité entre les paraphrases du corpus obtenu.
« A Monolingual Parallel Corpus of Arabic » [33] est un corpus de paraphrases, il contient plus de 100,000 de paires de paraphrases. Il est conçu pour l'arabe et généré automatiquement à partir de la traduction d'un corpus parallèle anglais-français à l'aide de l'API Google Translate et il est évalué par deux experts dans la langue arabe. Il s'agit du premier corpus monolingue parallèle d'arabe qui peut être utilisé pour entraîner des modèles séquence à séquences pour la paraphrase.

c. Extraction de corpus à partir de connaissance web

[5]Terme frequency & inverse document frequency : méthode d'alignement de documents.
[6]http://www.ldc.upenn.edu/

Cette technique basée sur la collection de données à l'aide de connaissance Web, principalement à partir de sites Web d'actualités, pour construire un corpus parallèle. L'utilisation du Web comme base pour la constitution de ressources textuelles est très récente. Ces dernières années ont connu des travaux tentant d'exploiter ce type de données.
Le corpus anglais-arabe [34] est collecté en utilisant le web mining, les pages de langues souhaitées sont téléchargées à partir de domaines spécifiés. Les fausses pages sont rejetées afin de créer un ensemble de paires candidates. Une correspondance basée sur le contenu est effectuée pour calculer la similitude de parallélisme entre chaque paire candidate à l'aide d'un dictionnaire anglais-arabe pour déterminer s'il s'agit d'une correspondance ou non.

Dans [35], la construction d'un grand corpus de paraphrases pour l'anglais et le néerlandais se fait en alignant des titres groupés qui sont extraits d'un site d'agrégation d'actualités. Les phrases obtenues sont évaluées en utilisant des jugements humains recueillis auprès de 76 participants. Pour générer des paraphrases sententielles, ils ont utilisé un cadre de traduction automatique basé sur des phrases (PBMT)[7] standard modifié avec un composant de re-classement (désormais PBMT-R).

L'article [36] traite des techniques non supervisées pour acquérir des paraphrases monolingues au niveau de la phrase à partir d'un corpus d'articles de presse regroupés dans le temps et par sujet et collectés à partir de milliers de sources d'information sur le Web. Le corpus obtenu appelé Microsoft Research Paraphrase Corpus (MSRP) contient 5 801 paires de phrases construit en utilisant deux techniques : (1) une simple distance d'édition de chaîne (levenshtein) qui compte le nombre de suppressions et d'insertions lexicales nécessaires pour transformer une chaîne en une autre, et (2) une stratégie heuristique qui associe des phrases initiales de différentes nouvelles dans le même groupe, elle est appuyée sur une heuristique basée sur le discours, spécifique au genre de l'information, pour identifier les paires de paraphrases probables même lorsqu'elles ont peu de similitudes superficielles.

4. Analyse et discussion

La plupart des travaux antérieurs sur des textes parallèles ont été menés sur quelques corpus parallèles construits manuellement tels que le corpus du hansard canadien[8] et le Consortium

7 La traduction automatique basée sur des phrases (PBMT : Phrase-Based Machine Translation) est une forme de SMT où le modèle de traduction vise à traduire des séquences de mots plus longues (« phrases ») en une seule fois.

8 hansard canadien basé sur les débats du Parlement canadien.

de données linguistiques (LDC). Cependant, la collecte manuelle de grands corpus est une tâche fastidieuse, longue et gourmande en ressources.

Malheureusement, les données parallèles librement disponibles sont aussi des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de taille raisonnable sont disponibles comme l'anglais, le français, l'espagnol, et quelques langues européennes. De plus, ces corpus proviennent principalement de sources gouvernementales, comme le parlement canadien ou européen, ou de l'Organisation des Nations Unies. Ceci est problématique, parce que les systèmes en TAL entraînés sur des données provenant, par exemple, d'un domaine politique ne donnent pas de bons résultats lorsqu'ils sont utilisés pour traiter des articles scientifiques par exemple.

Pour la langue arabe, et bien qu'elle soit riche en terme de vocabulaire et de linguistique, elle souffre de manque d'investissement en recherche. En effet, il n'y a pas de grands travaux sur la construction de corpus de paraphrases qui aident à consolider les recherches dans le domaine de TAL. À notre connaissance, le seul corpus de paraphrases dédié pour la langue arabe est celui de [34], son contenu comporte plusieurs erreurs et anomalies donc nécessite beaucoup de prétraitement avant qu'il soit utilisable, et n'est pas un corpus totalement bien formé par rapport (1) **la taille** de corpus qui contient presque 100,000 de paires de paraphrases, il est insuffisant pour couvrir une grande partie de la langue arabe qui contient plus de 12 millions de mots où il peut produire des milliards de phrases (avec une grande probabilité qu'elles forment des grandes classes liées par une relation de paraphrase), et (2) **le temps** couvert par les textes du corpus, car ce corpus c'est une traduction du corpus bilingue EuroParl-v7[9] qui contient des textes rédigés à des intervalles de temps trop larges qui présentent le risque d'être non pertinents. De plus, ce dataset contient des données de domaines générales, donc n'est pas applicable pour des traitements spécifiques. Il est impossible pour une seule traduction de référence de saisir toutes les verbalisations possibles pouvant véhiculer le même contenu sémantique. Cela peut pénaliser injustement les hypothèses de traduction qui ont le même sens mais utilisent des n-grammes qui ne sont pas présents dans la référence. Par exemple, la sortie système S donnée n'aura pas un score élevé par rapport à la référence R même si elle véhicule exactement le même contenu sémantique.

D'ailleurs, Les textes résultants de la traduction risquent de n'être pas idéales car lors

[9]Europarl-v7: version évalué de corpus parallèle Europarl pour la Machine statique de Translation.

de la traduction des textes scientifiques, un traducteur qui ne connaît pas la traduction d'un terme peut, par exemple, paraphraser le terme avec d'autres mots, utiliser des mots moins représentatifs pour expliquer le terme, omettre le terme du texte traduit, etc. De plus, la traduction des documents d'une langue source vers une langue cible peut être coûteuse en termes de temps et d'argent [37]. Par contre, Les corpus basés sur les connaissances web représentent un groupe de corpus très variés, très larges et assez polyvalents, ce qui les rends très facile à exploiter dans le TAL arabe.

Le **tableau 1** suivant met en évidence les différences et les caractéristiques que nous avons pu identifier de chaque approche de construction de corpus mentionnée. Parmi ces critères nous citons :

- La taille de corpus obtenu pour chaque approche et le domaine d'application ;
- La disponibilité et l'accessibilité de données exploitant l'approche ;
- Ainsi le cout de création de corpus en termes de temps.

Caractéristiques / Approches	Taille	Disponibilité de données	Domaine d'utilisation	Accessibilité des données	Coût En termes De temps (1)
Corpus parallèle à partir d'un corpus comparable	Limité	Disponible	Spécialisé et précis à un contexte	Pas toujours	Coûteuse
Corpus parallèle à partir d'un corpus bilingue	Limité	Disponible pour Certaines langues	Spécialisé et précis à un contexte	Pas toujours	Coûteuse
Corpus parallèle à partir de connaissances web	Non limité et Varié selon Le domaine	Disponible largement	Spécialisé, Multi-domaine, général	Accessible en grande quantité	Non coûteuse

Tableau 1 : tableau représente une comparaison entre les différentes approches d'extraction de corpus parallèle.

(1) Le coût en termes de temps dépend de la méthode de construction utilisée, la création d'un corpus parallèle à partir d'un corpus comparable ou d'un corpus bilingue nécessite que les traductions soient souvent effectuées à l'aide de traducteurs humains, ce qui consomme

certains temps et parfois l'argent. Tandis que l'exploitation de web pour la collection de données ne nécessite que des machines un peu puissantes.

Conclusion

La Paraphrase d'une phrase est une forme de surface alternative dans le même langage exprimant le même contenu sémantique que la forme originale de la phrase. Les corpus de paraphrases à large échelle sont importants dans de nombreuses applications de TAL. Dans cette partie, nous avons fait une étude bibliographique sur les différentes notions liées à notre contexte ainsi les travaux connexes existants.

Les étapes de création de corpus de paraphrases monolingues fait l'objectif de **chapitre 2** dans lequel nous allons entamer le processus de construction du corpus de paraphrases en se basant sur la collection des données à partir du web et l'application des techniques d'alignement des paires de phrases. Ainsi, nous allons présenter les méthodes d'évaluation.

Chapitre 2 : La Création de corpus monolingue de paraphrases

1. Introduction

Les corpus parallèles de paraphrases sont une ressource précieuse pour différentes applications de TAL, mais à l'heure actuelle, leur disponibilité et leur utilité sont limitées, les restrictions de licence et la difficulté fondamentale de localiser des textes parallèles dans toutes les langues du monde, sauf la plus dominante. Une ressource de corpus parallèle non encore largement explorée est le World Wide Web, qui héberge une abondance de pages parallèles, offrant une solution potentielle au problème de rareté des corpus de paraphrases. Une façon de pallier ce manque de données parallèles est d'utiliser une approche basée sur les connaissances web pour l'extraction de corpus parallèles.

Dans ce chapitre, nous allons procéder à présenter :

- La méthodologie de Conception relative à notre approche,
- Les choix de réalisation de création de corpus de paraphrases, et
- L'évaluation quantitative et qualitative du corpus construit et de l'approche correspondante.

2. Méthodologie de Conception

La motivation fondamentale de la modélisation est de fournir une démarche antérieure afin de réduire la complexité du système étudié lors de la conception et d'organiser la réalisation du projet en définissant les modules et les étapes de la réalisation.

Dans notre travail, nous nous concentrons sur l'extraction de documents et de phrases parallèles monolingues pour construire un grand corpus d'apprentissage afin d'entrainer un système de génération automatique de paraphrases pour la langue arabe. La méthode d'extraction de corpus parallèle à partir de connaissances du Web est appliquée. L'apport essentiel ajouté par notre approche est qu'elle permet de recueillir les données de différents sites en utilisant des mots clés pour faciliter la recherche, puis utiliser une méthode d'alignement basée sur l'aspect sémantique et syntaxique des phrases. Tandis que les autres

travaux ont basés sur la collection de données depuis un site spécifique, ils utilisent souvent des mesures syntaxiques seulement pour aligner les phrases.

Dans notre méthode, le système, tout d'abord, va analyser la requête effectuée afin d'obtenir des pages Web de domaine cherché, puis va créer des paires de documents parallèles candidates. Enfin, une technique d'extraction et de mesure de la similitude entre les paires de phrases candidates sera appliquée. La procédure générale de constitution de notre corpus est illustrée dans la **figure 3** suivante et divisée en trois grandes étapes :

1) la collection de données depuis le web,

2) l'extraction de phrases,

3) l'alignement de phrases obtenues.

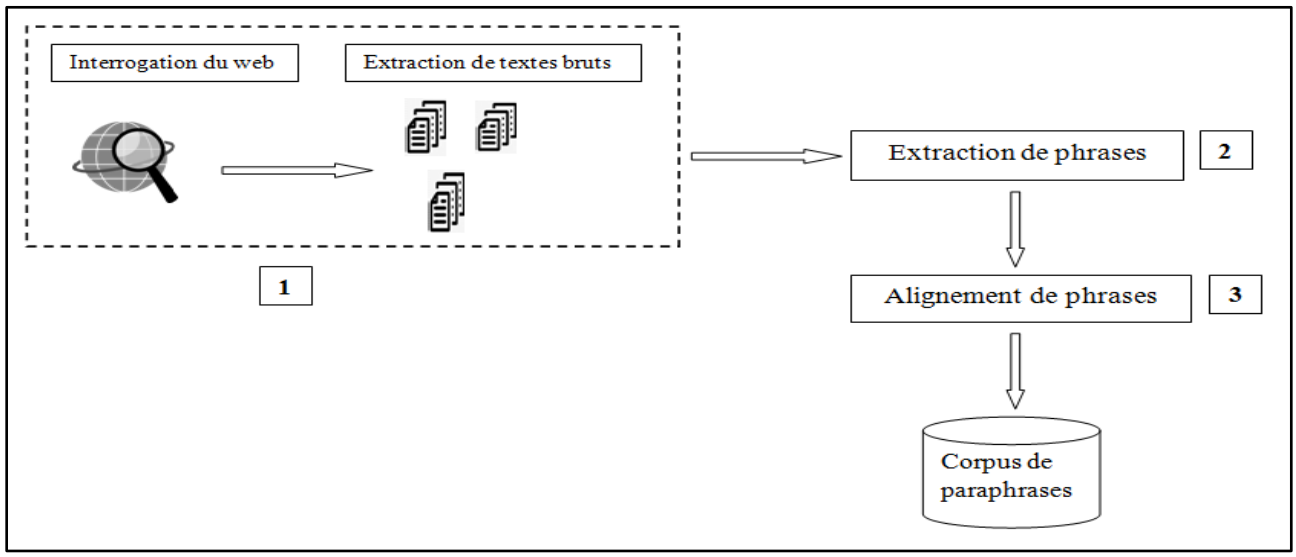


Figure 3 : la procédure générale de construction de corpus à partir du web.

2.1. La collection de données depuis le web

L'objectif principal de ce travail est de fournir un grand corpus de paraphrases pour entraîner les modèles de génération de paraphrases et assurer d'autres tâches de TAL dans différents domaines de texte. À cette fin, nous avons décidé de collecter des données pour construire notre dataset à partir de deux sources de données qui sont : le World Wide Web et le Wikipedia. La justification de la collecte de ces données ainsi que les mesures prises pour collecter les données de chacune de ces ressources sont détaillées dans les sous-sections suivantes.

2.1.1. L'interrogation de moteur de recherche

L'arabe est la cinquième langue la plus utilisée au monde avec plus de 420 millions de locuteurs natifs. Plus de 41 % d'arabophones utilisent Internet, ce qui représente environ 4 %

du contexte Internet à la fin de 2016. Les sites Web publics trouvés sur le World Wide Web couvrent un large éventail de sujets répartis sur des sites d'actualités, des blogs, des sites de services et des forums sociaux, et beaucoup d'autres. Pour collecter les données depuis le web, il est possible d'utiliser de nombreux moteurs de recherche comme des outils de recherche d'informations.

- **Le moteur de recherche Google**

C'est un moteur de recherche gratuit et libre d'accès sur le World Wide Web, ayant donné son nom à la société Google. C'est aujourd'hui le moteur de recherche et le site web le plus visité au monde.

Il peut être considéré comme étant une source de données linguistiques, pour extraire les passages pertinents qui sont susceptibles de contenir la réponse précise à une question donnée ou à la consolidation de corpus qui permettrait de le faire.

- **Wikipedia**

Wikipedia est une encyclopédie numérique, multilingue, libre et universelle officiellement née en janvier 2001. L'édition et la publication des textes sont réalisées grâce à un logiciel appelé moteur de wiki. La modification ou l'ajout des textes se fait en ligne directement dans le navigateur Web.

Chaque article de Wikipédia est référencé de manière unique par une adresse URL ; ce qui élimine tout risque d'ambiguïté. Utiliser Wikipédia comme corpus parallèle présente plusieurs avantages.

- la masse de données est intéressante.
- La variété des thèmes abordés permet de constituer un corpus à la fois général et spécialisé.
- Tout le matériel est utilisable librement car l'ensemble de l'encyclopédie est sous licence libre Créative Commun (*Creative Commons).*

L'interrogation de moteur de recherche est faite en effectuant une requête avec des mots clés concernant le domaine recherché afin de collecter une liste des pages web contenant les données nécessaires pour la construction de notre corpus.

2.1.2. Extraction des textes bruts

Cette étape nous permet de nettoyer les fichiers html et extraire les textes bruts c'est-à-dire d'enlever toutes les balises du fichier html de façon à ne garder que le texte.

2.2. Extraction de phrases

Dans cette phase, les textes bruts collectés de l'étape précédente sont découpés en phrases,

cette étape est appelée «Tokenisation ».

La tokenisation : signifie la segmentation du texte selon un caractère identifié (généralement le point de fin de phrases, le point d'interrogation…...). Elle aide à interpréter la signification du texte en analysant la séquence des mots et aide à comprendre le contexte d'étude. L'utilité de tokenisation dans notre travail est d'extraire les phrases nécessaires à utiliser après pour construire les paires de phrases.

Cependant, la collecte de données depuis le web provient de nombreuses sources. Cela nécessite une étape de nettoyage de texte non arabe.

Filtrage du contenu non arabe : filtrer la collection de données de contenu non arabe est particulièrement important lorsqu'il s'agit de données provenant du « Web ». L'arabe peut être facilement reconnu à l'aide de son alphabet ("ت ,ب ,أ ,..."). Cette étape doit précéder la segmentation pour assurer que les phrases sont nettement arabes.

Les étapes d'extraction de phrases sont illustrées dans la **figure 4** :

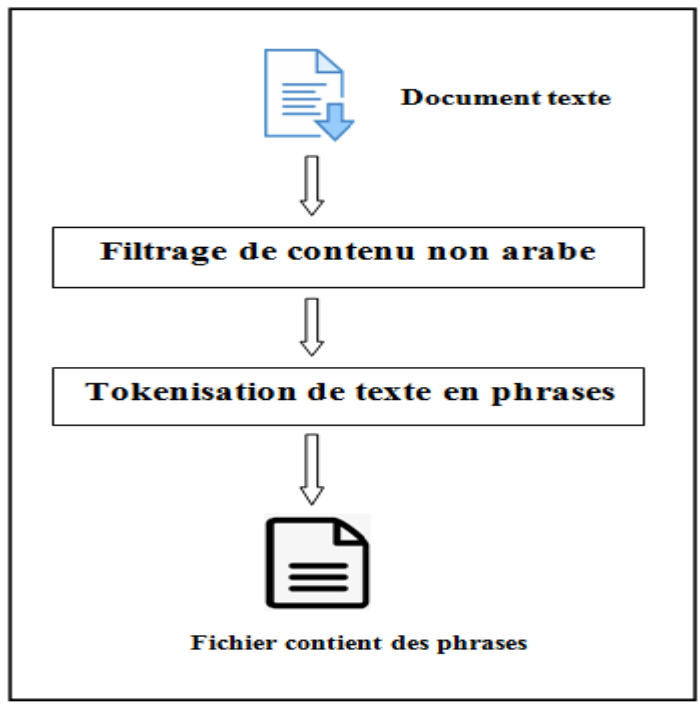


Figure 4 : Extraction de phrase.

2.3. L'alignement des phrases

L'étape d'alignement des phrases sert à mettre en correspondance les phrases qui sont en relation de paraphrase. Les paires de paraphrases peuvent être identifiées en calculant diverses mesures de similarité/distance de phrases entre les deux phrases d'une paire.

L'alignement des phrases est un domaine de recherche très actif, la tâche d'aligner des

phrases parallèles a reçu une attention considérable depuis la renaissance de la traduction automatique basée sur les données.

L'alignement est un processus qui consiste à comparer deux textes et déterminer les segments correspondants.

Les méthodes d'alignement automatique des phrases sont généralement confrontées à deux types de difficultés. 1) il y a la question de la robustesse. En réalité, les divergences entre un texte source et un texte cible sont assez courantes : différences de forme, inversions, etc. Les programmes d'alignement de phrases doivent être prêts à faire face à de tels phénomènes. 2) il y a la question de l'exactitude, certaines décisions sont difficiles à prendre même pour les humains.

Pour produire un programme d'alignement des phrases à la fois robuste et précis, nous avons combiné deux méthodes : l'une qui mesure la similitude au niveau syntaxique, et l'autre sémantique qui s'appuie sur le sens de mots et le contexte.

2.3.1. Normalisation de données

Avant d'effectuer tout type d'alignement, le corpus nécessitent une étape de normalisation qui permet d'homogène les données. Parmi telles traitements nous citons les tâches suivantes:

- **La normalisation :** la normalisation des caractères est une étape de prétraitement courante lorsqu'il s'agit de texte arabe. Il permet de ne pas prendre en compte des détails importants au niveau local (ponctuation, les signes diacritiques,… etc.)
- **La suppression des mots vides** ou **stop-words**, les mots inutiles (…, و, مع, إلى), car les mots identiques considérés comme peu pertinents peuvent parfois trop influer sur la valeur de la similarité. Par exemple : « تاريخ الاقتصاد في العالم» et « الاقتصاد في قارة إفريقيا», le terme "في" n'est pas vraiment pertinent mais il va avoir un poids certain qui influence sur le résultat, donc, l'élimination des mots-vides permettent de pallier ce problème. Ils sont supprimés du vocabulaire pour réduire le bruit dans le corpus.
- **Le stemming :** est le processus de production de variantes morphologiques d'un mot racine/base de la langue [38]. Le mot racine est appelé lemme. Un algorithme de lemmatisation réduit les mots « معلومات » à la racine du mot « علم ».

Dans le cadre de notre projet, nous avons implémenté deux méthodes de stemming à savoir :

- **Tashaphyne**[10] : est une bibliothèque python. Il s'agit d'un stemmer léger et un segmenteur arabe. Il supporte principalement le light stemming (suppression des préfixes et suffixes) et donne toutes les segmentations possibles. Il offre en même temps l'extraction des racines (root stemming ou stemming lourd).
- **ISRI Arabic Stemmer**[11] **:** est l'un des stemmers arabes contenus dans le package NLTK (nltk.stem), il renvoie les mots racines en utilisant des modèles morphologiques des mots arabe.

2.3.2. Les approches de mesure de similarité utilisées

Une mesure de similarité est, en général, une fonction qui quantifie le rapport entre deux objets, comparés en fonction de leurs points de ressemblance et de dissemblance.

Dans la littérature, différentes mesures de similarité sont proposées [39]. Certaines approches couramment utilisées exploitent la structure syntaxique des phrases ; le nombre de tokens ou n-grams en commun entre la phrase source et la phrase cible est généralement calculé. D'autres tentent de prendre en compte les problèmes de synonymie et la sémantique des phrases en exploitant des ressources sémantiques ou des méthodes statistiques [40].

Nous avons combiné des mesures comme l'illustre **la figure 6**.

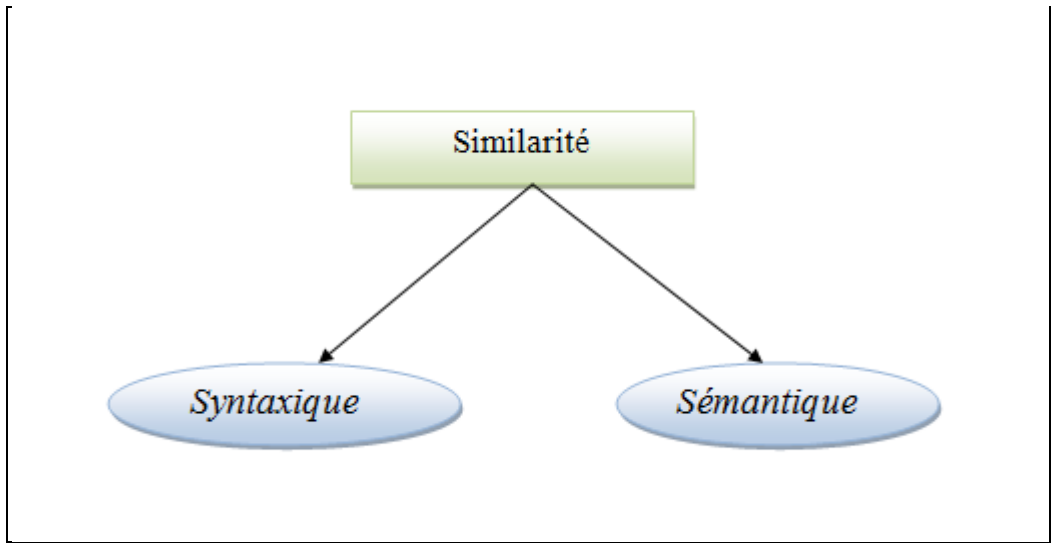


Figure 5 : les types de similarités utilisés.

[10]Disponible sur https://pypi.org/project/Tashaphyne/0.3/

[11]Dispnible sur https://www.nltk.org/_modules/nltk/stem/isri.html

2.3.2.1. Similarité syntaxique

C'est une métrique qui mesure la similarité ou la dissimilarité entre deux chaînes de caractères au niveau morphologique c-à-d prendre en considération la structure de la chaîne.

Parmi de telles mesures de similarité, les mesures basées sur les termes (mesure de Dice, Jaccard, la distance euclidienne), les mesures utilisant les séquences de caractères (la distance d'édition de Levenshtein, Jaro….).

- **Les mesures basées sur les caractères utilisées**

Levenshtein distance : elle a été proposée par Vladimir Levenshtein en 1965. Elle est également connue sous les noms de distance d'édition ou de déformation dynamique temporelle[41]**.**

C'est une distance donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

Chaque modification a un coût égal à 1.

$$LDN = \frac{N}{max\ (n,\ m)}$$

Équation 1 : formule de distance de Levenshtein (LDN).

Où N est le nombre minimum d'opérations de chaîne à transformer un texte x en y ou vice versa, et n et m sont les nombre de symboles dans les textes x et y respectivement.

Exemple

Si P = « examen » et Q = « examen », alors LDN (P, Q) = 0, parce qu'aucune opération n'a été réalisée.

Si P = « exam**e**n » et Q = « exam**a**n », alors LDN (P, Q) = 1, parce qu'il y a eu une substitution (changement du e en a), et que l'on ne peut pas en faire moins.

****Ladistance de Jaro***: mesure la similarité entre deux chaînes de caractères[42]. Elle est principalement utilisée dans la détection de doublons. Le résultat est normalisé de façon à avoir une mesure entre 0 et 1, donc zéro représente l'absence de similarité et 1, l'égalité des chaînes comparées.

La distance de Jaro entre chaînes S_1 et S_2 est définie par :

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right)$$

Équation 2 : formule de jaro.

Où:

- S_i est la longueur de la chaîne de caractères ;
- m est le nombre de caractères correspondants (1) ;
- t est le nombre de transpositions (2).

(1) Deux caractères identiques de S_1 et S_2 de sont considérés comme correspondants si leur éloignement (i.e. la différence entre leurs positions dans leurs chaînes respectives) ne dépasse pas :

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1.$$

(2) Le nombre de transpositions est obtenu en comparant le $i^{ème}$ caractère correspondant de S_1 avec le $i^{ème}$ caractère correspondant de S_2. Le nombre de fois où ces caractères sont différents, divisé par deux, donne le nombre de transpositions.

➢ Les mesures basées sur les termes

****La distance euclidienne (ED) :*** pour deux vecteurs v et u de taille N, la distance ED est exprimée comme étant :

$$dED(v,u) = \sqrt{\sum_{i=1}^{n} (ui - vi)^2}$$

Équation 3 : distance euclidienne.

Où n est le nombre total de termes représentés, i.e. la taille des vecteurs.

****Similarité Jaccard*:** L'indice de Jaccard calcule la similarité entre deux chaînes de caractères en se basant sur les tailles des ensembles « union » et « intersection » de deux séquences S1 et S2 [43].

$$JaccSim = || Intersection (S1, S2) || / || Union (S1, S2) ||$$

Équation 4 : formule de Jaccard.

Exemple

S1 = L'indice de Jaccard calcule la similarité entre deux chaînes de caractères.

S2 = L'indice de Jaccard utilise les tailles des ensembles pour mesurer la similarité.

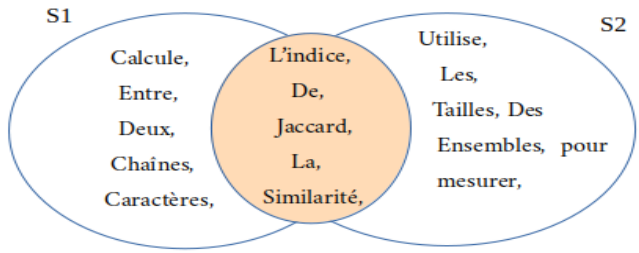


Figure 6 : principe de méthode Jaccard.

***Indice de Dice* :** mesure la similarité entre deux phrases P_1 et P_2 en se basant sur le nombre de termes communs à P_1 et P_2 [44].

$$SimDice(p1, p2) = \frac{2N_c}{N_1+N_2}$$

Équation 5 : formule de Dice.

Où : N_c est le nombre de termes communs à P_1 et P_2, et N_1 (resp. N_2) est le nombre de termes de P_1 (resp. P_2).

Le problème qui se pose lors de l'utilisation des techniques basées sur l'approche syntaxique est qu'elles ne prennent pas en compte la sémantique. Par exemple : il est difficile de trouver une forte similarité entre "Je possède un chien" et "J'ai un animal". Donc, la prise en compte de la sémantique semble importante.

2.3.2.2. Similarité sémantique et utilisation de Word Embeddings

Une mesure de similarité sémantique est un concept selon lequel un ensemble de documents ou de termes se voient attribuer une métrique basée sur la ressemblance de leur signification/contenu sémantique.

L'une des étapes majeures de tout modèle de TAL consiste à transformer les données textuelles de leur nature non structurée en une représentation structurée pour être prêtes pour le traitement [45]. Les machines nécessitent que les données soient converties dans un format numérique pour effectuer toute tâche d'apprentissage automatique. Afin d'effectuer de telles tâches, diverses techniques d'intégration de mots sont utilisées, (i.e Bag of Words[12], TF-IDF, word2vec) pour coder les données textuelles. Cela permet d'effectuer des opérations TAL telles que trouver une similitude entre deux phrases pour extraire des couples sémantiquement similaires du corpus (comme le cas de Frequently Asked QuestionsFAQ[13]), rechercher des documents similaires dans la base de données, recommander des articles de presse sémantiquement similaires…..

➔ Représentation des mots et phrases par les Word Embedding

Le word embedding « plongement de mots » également connue sous le nom de représentation de mots, joue un rôle important dans la construction de vecteurs de mots continus basés sur leurs contextes dans un large corpus. C'est une méthode d'apprentissage utilisée notamment en traitement automatique des langues. Cette technique permet de représenter chaque mot d'un dictionnaire par un vecteur de nombres réels. Le plongement de mots capture à la fois les informations sémantiques et syntaxiques des mots et peut être utilisé pour mesurer les similitudes de mots, qui sont largement utilisées dans diverses tâches IR et NLP.

Le WE consiste à construire des vecteurs de taille fixe qui prennent en compte le contexte dans lequel se trouvent les mots. Ainsi, deux mots présents dans des contextes similaires auront des vecteurs plus proches (en termes de distance vectorielle). Cela permet alors de capturer à la fois des similarités sémantiques, syntaxiques ou thématiques des mots.

➔ *Word2Vec :*

Word2vec est un réseau neuronal[14] à deux couches qui traite le texte en « vectorisant » les mots. Son entrée est un corpus de texte et sa sortie est un ensemble de vecteurs qui représentent les mots de ce corpus. Le but et l'utilité de Word2vec est de regrouper les

[12]Le modèle du sac de mots est une représentation simplificatrice utilisée dans le TAL et RI. Dans ce modèle, un texte (phrase, document) est représenté comme le sac (multiset) de ses mots, sans tenir compte de la grammaire et même de l'ordre des mots mais en gardant la multiplicité.

13 La **FAQ** (*Frequently Asked Questions*), également nommée « Foire Aux Questions », est une liste de réponses aux questions les plus fréquemment posées**.**

14 Le réseau de neurones se sert d'un algorithme pour apprendre des données nouvelles à partir d'exemples préalablement enregistrés et qu'il analyse rigoureusement. Il s'agit là d'un véritable mode d'apprentissage par l'expérience.

vecteurs de mots similaires dans un espace vectoriel qu'il détecte mathématiquement les similitudes. Word2vec crée des vecteurs qui sont des représentations numériques distribuées des caractéristiques des mots telles que le contexte de mots individuels. Cette approche est la plus couramment utilisée de nos jours dans la construction de modèles d'intégration de mots pour les différentes tâches basées sur le TAL en raison de leur performance supérieure dans la découverte des relations syntaxiques et sémantiques entre les mots dans l'espace continu.

L'une des variantes de Word2vec la plus utilisé est le modèle de Skip-Gram, c'est un framework connu pour l'apprentissage des vecteurs de mots, comme le montre la **figure 8**. Skip-Gram vise à prédire les mots de contexte donnés à un mot cible dans une fenêtre glissante. Dans ce cadre, chaque mot correspond à un vecteur unique. Le vecteur du mot cible est utilisé comme caractéristiques pour prédire les mots de contexte.

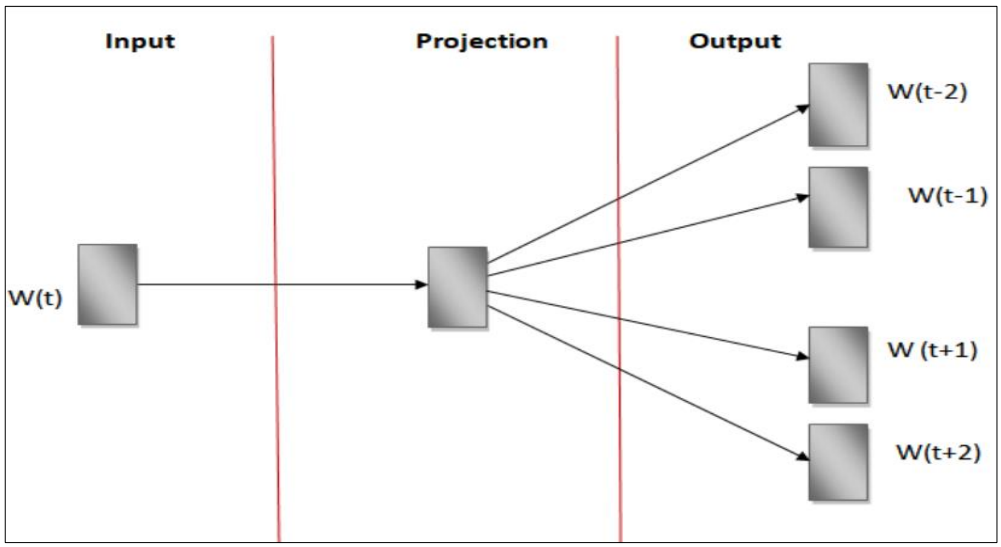


Figure 7 : l'architecture de modèle Skip-Gram [46].

-La couche d'entrée correspond au mot cible ;

- La couche de sortie correspond au contexte à prédire (target) à partir du mot en entrée.

Le but du réseau de neurones SkipGram est de maximiser l'équation suivante [46] :

$$\frac{1}{V}\sum_{t=1}^{V}\sum_{j=t-c,j\neq t}^{t+c}\log p(m_j \backslash m_t)$$

Où V (et C) est la taille du vocabulaire (contexte).

Le **tableau 2**suivant représente quelques caractéristiques du modèle Skip-Gram

	Temps d'entraînement de word2vec	Difficulté de Tâche d'entraînement	La sensibilité aux mots rares et fréquents (1)	Performance
Skip-Gram	Jusqu'à 3 jours en moyen	Capturer De Meilleures relations sémantiques	moins sensible	Plus performant et efficace

Tableau 2 : caractéristiques de modèle Skip-Gram.

(1) Étant donné que Skip-gram repose sur la saisie de mots simples, il est moins sensible aux mots fréquents surdimensionnés, car même si les mots fréquents sont présentés plus de fois que les mots rares pendant la formation, ils apparaissent toujours individuellement, ce qui conduit Skip-gram à être également plus efficace en termes de documents requis pour atteindre de bonnes performances [37].

L'un des travaux le plus important sur la construction des représentations des mots qui prend en charge de la langue arabe en utilisant Skip-Gram est celui de Zahran [47] qui a proposé un grand dictionnaire contenant les WE arabes. Le modèle qu'il a présenté a été approuvé son efficacité quantitativement et qualitativement.

Dans cette partie, nous avons introduit l'aspect général de notre démarche, nous avons capturé toutes les méthodes s'avèrent importantes pour procéder à réaliser concrètement notre travail à savoir la construction du corpus.

3. Réalisation

L'apport essentiel de ce travail consiste à proposer une approche méthodologique pour la création d'un corpus monolingue parallèle de paraphrases en se basant sur les données hébergées dans le web.

Pour implémenter notre approche, nous avons choisi le langage de programmation python qui nous semble être le mieux adapté à notre recherche. En effet, Python s'annonce comme une des évolutions majeures de la programmation et de l'intelligence artificielle. Pour la première fois, un langage efficace, performant, standard et facile à apprendre (et, de plus, gratuit) est disponible. Elle présente une très grande bibliothèque, variée et robuste pour plusieurs types de développement y compris le domaine du TAL.

Notre démarche d'implémentation repose sur les points qui sont introduits dans la partie 'méthodologie de conception' à savoir : l'extraction de données depuis le web, l'application de différentes fonction de prétraitement de données (normalisation, élimination de mots inutiles, ….) et l'alignement des phrases obtenues. Les étapes à suivre lors de cette phase sont les suivantes :

3.1. Etape1 : La collection de documents parallèles candidats

Le but de cette étape est de construire la liste des fichiers html obtenues suivant la recherche effectuée sur Google et Wikipedia, cette procédure appelé le Web Scraping. Pour ce faire, nous avons implémenté un script Python qui prend en entrée un ensemble de mots clés et génère en sortie l'ensemble des fichiers html correspondants et pour cela nous avons utilisé la bibliothèque« request[15] » de python pour récupérer les données à analyser. Nous avons ainsi offert une possibilité de limiter le nombre de pages à retourner pour chaque recherche en paramétrant notre programme afin de faciliter le processus de collection.

3.2. Etape2 : Extraction de textes bruts

Comme nous avons indiqué précédemment, cette étape sert à nettoyer de textes de balises html. Pour effectuer cette tache, nous avons utilisé Beautiful Soup[16] qui est une bibliothèque Python permettant d'extraire des données de fichiers HTML et XML.

3.3. Etape3 : Obtention des phrases

Cette étape consiste à extraire les phrases à partir les textes déjà préparer dans l'étape 2.

Pour en arriver là, nous avons essayé d'extraire les phrases qui représentent la donnée élémentaire de notre travail, pour cela nous avons utilisé des outils pour découper le texte en ensemble de tokens appelées ici phrases.

Python offre une bibliothèque de tokenisation appelée NLTK[17] avec une fonction sent_tokenize() qui appartienne au module « nltk.tokenize.punkt », qui a déjà été formée et sait donc très bien marquer la fin et le début de la phrase à quels caractères et ponctuation. Cette fonction prend en entrée un texte et rend un ensemble de token en sortie (sont les phrases à traitées).

Une fonction d'élimination de tout contenu non arabe s'avère nécessaire dans cette étape, ce processus est fait d'une façon à ne garder que les mots arabes pour ne pas tomber dans le problème de reconnaissance de langage qui influence sur les résultats de similarité par la

[15] Disponible sur https://fr.python-requests.org/en/latest/user/install.html#install

[16] Disponible surhttps://pypi.org/project/beautifulsoup4/

17 NLTK est une plate-forme utilisée pour créer des programmes Python pour traiter des données sous forme de langage.

suite et éviter tout bruit qui peut être induite par ça.

Les résultats sont conservés dans des fichiers textes (.txt), ou chaque ligne contient une phrase, pour nous permettre de les utiliser dans l'étape prochaine d'alignement des paires de phrases.

Jusqu'ici, Nous en avons terminé avec la première étape vers la création de notre corpus de paraphrases, les étapes suivantes concernent la préparation des paires de phrases candidates.

3.4. Etape4 : Alignement de paires de phrases

Pour aligner notre corpus nous avons basé sur le processus représenté dans la **figure 8** suivante :

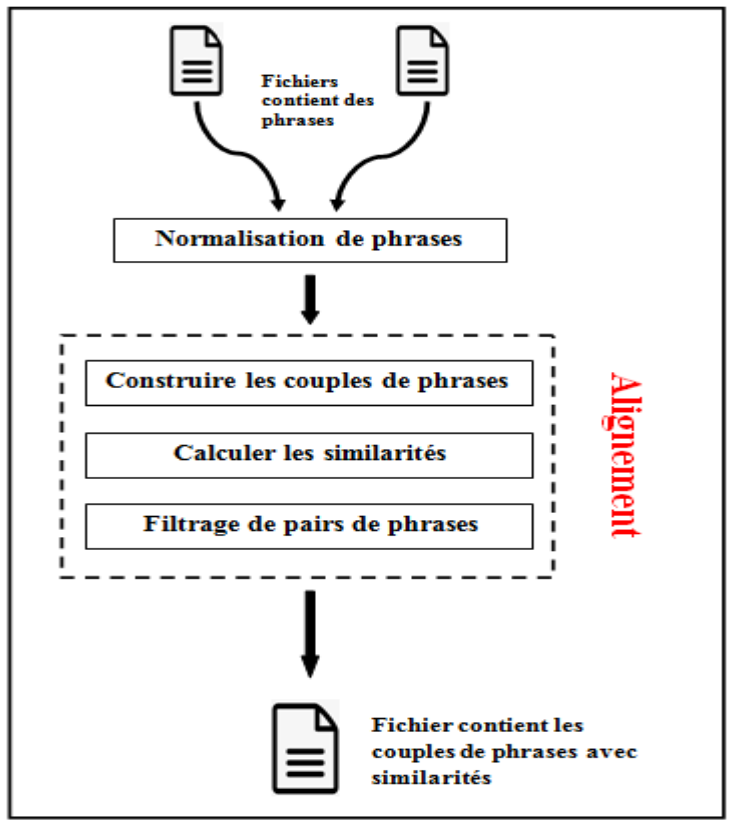


Figure 8 : Le processus d'alignement de phrases.

3.4.1. Normalisation de phrases

La normalisation est une technique qui permet de préparer les données reçues à la phase d'alignement. Nous avons éliminé toutes les données et méta données non nécessaires pour notre travail. Pour cela nous effectuons un certain nombre de tâches, afin de garantir la qualité des données constituées notre corpus :

- Normaliser les phrases pour simplifier la caractérisation de la forme (caractère, chiffre, mot..) et diminuer la quantité d'information à traiter. Dans cette étape, les

lettres « إ, أ et آ » sont remplacées par « ا » tandis que la lettre « ة » est remplacée par « ه » et la lettre « ي » est remplacée par «ى». Les signes diacritiques sont également supprimés dans cette étape. Ainsi, tous les caractères spéciaux (§, %, $, ….) sont supprimés.

Exemple :

Avant	Après
جريمة	جريمه
إقتصاد	اقتصاد
قَانُونْ	قانون

Tableau 3 : exemple de mots avant et après la normalisation.

- Eliminer les mots non utiles ou encore appelés les Stop Words. Il existe dans la librairie NLTK une liste par défaut des stopwords dans plusieurs langues, notamment l'arabe. Mais nous fait ceci d'une autre manière: nous avons récupéré une grande liste de stops words d'ici[18] , puis nous avons supprimé les mots les plus fréquents du phrase et considérer qu'ils font partie du vocabulaire commun et n'apportent aucune information.
- La tâche de stemming des mots nécessite encore une fois de ne conserver que le sens des mots utilisés dans le corpus. Cela consiste à ne conserver que la racine des mots étudiés. L'idée étant de supprimer les suffixes, préfixes et autres des mots afin de ne conserver que leur origine. En effet, nous l'avons met comme une tache supplémentaire peut être utilisé à la demande car notre travail repose sur l'aspect sémantique du mot donc laisser le mot à son état ne va pas influencer sur les résultats finaux. Les deux techniques expliquées précédemment sont utilisées ; ISRI-Stemmer et Tashapyne avec ses deux variantes, Tasha Light Stemmer et Tasha Root Stemmer. Des exemples d'application sont présentés dans le tableau suivant :

[18] https://github.com/mohataher/arabic-stop-words/blob/master/list.txt

Mot	ISRI Stemmer	Tasha lite stemmer	Tasha root stemmer
فروع	فرع	فرع	فرع
المعروفة	عرف	معروف	عرف
المعلومات	علم	معلوم	علم
استحداث	حدث	استحداث	حدث
معاملات	عمل	معامل	عمل

Tableau 4 : exemple de mots stemmés.

On remarque que la lemmatisation en utilisant ISRI et Tashapyne change le sens de quelque mot dû aux les stemmers qui existent pour la langue arabe ne présentent pas de documentation disponible et ne présentent pas une évaluation de la précision des résultats obtenus.

- Délimiter la taille de phrase en supprimant toute phrase jugée trop courte (moins de 3 mots) ou trop longue (plus de 50 mots).Cette contrainte a diminué la taille du corpus, mais nous a permis quand même d'avoir un corpus plus homogène.

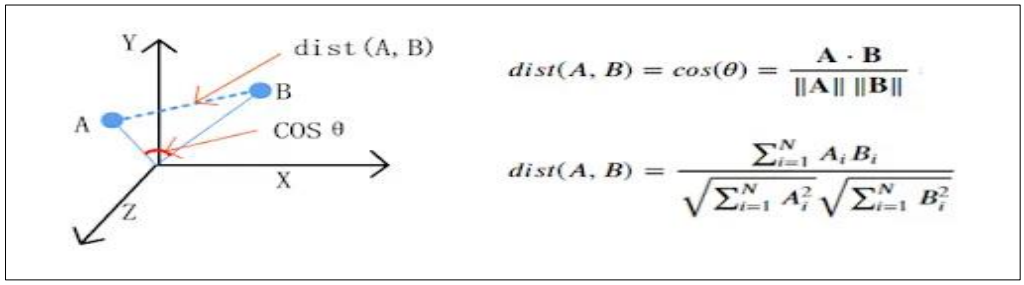
3.4.2. Construire les couples de phrases

Au cours de cette phase, nous avons couplé les fichiers deux à deux de tel sorte qu'une phrase de fichier A est couplée avec toutes les phrases de fichier B. Toutes les paires de phrases résultantes ont été considérées comme des paires de paraphrases candidates.

3.4.3. Calculer les similarités

Pour aligner les phrases et garantir la qualité de la relation sémantique-syntaxique des mots, nous avons proposé de combiner la méthode syntaxique (i.e. tenir compte du changement des caractères entre les chaînes à savoir la méthode de jaccard, jaro, dice, levenshtein et la distance euclidienne) et de modèle sémantique intégrant le sens des mots dans un contexte en utilisant l'architectures Skip-gram de la méthode Word2vec en utilisant le dictionnaire de word embedding de Zahran[47] pour la langue arabe pour désigner les sens des mots. La taille du vecteur de WE utilisé est égale à 300 composants.

Nous avons utilisé la métrique Cosine et Cosine combiné avec Tf-minmax. Pour chaque couple C, nous créons une représentation avec WE pour (c1) et (c2), et calculons la similarité de cosinus entre (c1) et (c2). Plus la similarité Cosine est proche de 1, plus les deux mots sont liés.



- **Tf-Max** est utilisé pour supprimer les termes qui apparaissent trop fréquemment, également appelés "mots vides spécifiques à un corpus"
- **Tf-Min** est utilisé pour supprimer les termes qui apparaissent trop rarement.

$$TFminmax(w) = TFlog(W)/Max(TFlogs)$$

Figure 9 : la formule de Tf-MinMax.

Où :

- Tflog(w) : la fréquence d'apparition d'un mot w dans l'ensemble de données.
- Max(TFlogs) : nombre maximal dans le TFlogs.

La figure suivante représente un exemple de l'application de différentes méthodes de calcul de similarité syntaxique et sémantique sur un couple de phrases :

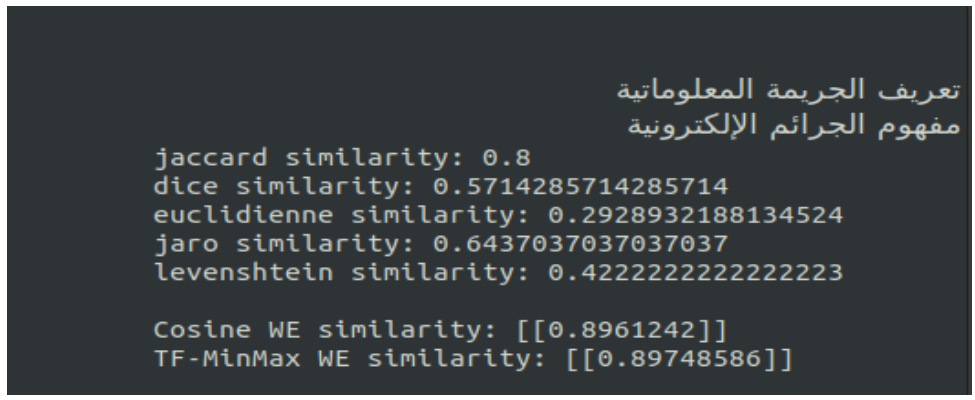


Figure 10 : exemple de résultat de similarité entre deux phrases en utilisant les différentes mesures d'alignement.

Nous remarquons que le résultat de similarité diffère considérablement soit entre l'approche syntaxique et l'approche sémantique (i.e. l'une repose sur la forme de mots et caractères,

tandis que l'autre se base sur l'aspect sémantique de mots) ou encore entre les métriques de même approche à cause de la manière traitant le mot de chaque méthode.

3.4.4. Filtrage des paires de phrases

Le filtrage de paires de phrases sert à regrouper les pairs de phrases en des classes selon le degré de similarité.

Après l'obtention des similarités entre les couples, un passage au score va être effectué et pour cela nous avons proposé deux méthodes :

1) Calculer le centre de similitude entre les résultats de calcul de similarité en calculant la moyenne arithmétique donnée par la formule (1) de mesures de similarité syntaxique (de même pour les mesures sémantiques) utilisées précédemment. Après, nous avons calculé la moyenne harmonique en utilisant la formule (2) entre les deux moyennes arithmétiques résultantes.

$$Moy_{arithm} = \Sigma \text{ mesures} / N \qquad (1)$$

Équation 6 : formule de la moyenne arithmétique.

Où N est le nombre de mesures calculées.

$$Moy_{harmo} = 2(M1 * M2)/ (M1 + M2) \qquad (2)$$

Équation 7 : formule de la moyenne harmonique.

Où :

- M1 (respectivement M2) est la moyenne arithmétique de mesures syntaxiques (respectivement sémantiques).

En appliquant ces calculs sur l'exemple précédent, les résultats seront les suivants :

```
Arithmitic AVG Syntaxic: 0.54604954323359
Arithmitic AVG Semantic: 0.8968050302330695
Harmonic AVG: 0.678793256276384
```

Figure 11 : exemple de la moyenne arithmétique et harmonique.

Nous avons remarqué que le taux de similarité entre les deux couples a diminué car les résultats syntaxiques ont influencé sur les calculs, même s'il y a une grande probabilité entre ces deux couples d'être similaires. Pour cela, nous avons procédé à utiliser la pondération des résultats.

2) Calculer la moyenne pondérée entre les couples en appliquant deux probabilités, la première prend une pondération (3) de 0.3 pour la l'approche syntaxique et 0.7 pour l'approche sémantique. La deuxième pondération (4) prend un score de 0.2 et 0 .8 (syntaxique, sémantique respectivement).
Nous avons obtenu ces valeurs de pondérations par essai successif sur de grands échantillons du corpus réalisé.

$$\textbf{Ponderation1} = Moy_{syntax} * 0.3 + Moy_{semantic} * 0.7 \quad (3)$$

$$\textbf{Ponderation2} = Moy_{syntax} * 0.2 + Moy_{semantic} * 0.8 \quad (4)$$

Figure 12 : méthode de calcul de moyenne pondérée.

Où :
- Moy_{syntax} représente la moyenne arithmétique syntaxique.
- $Moy_{semantic}$ représente la moyenne arithmétique sémantique.

Dans une première interprétation, nous avons remarqué que les résultats obtenus à partir de l'application des deux méthodes (la moyenne pondérée1 et la moyenne pondérée2) sont proches. Donc, le choix a été de classifier notre corpus par rapport à la deuxième pondération (0.2, 0.8) car elle rapproche bien la similarité entre les couples de phrases.
Les couples sont classés selon trois catégories :

- **Classe1 :** les paires de phrases ayant une similarité combinée supérieure à 0.80 partagent de maximum de détails entre eux sont considérées comme des paraphrases (voir des vraies paraphrases). Ces paires sont introduites directement dans le corpus pour entrainer un modèle de génération de paraphrases ;
- **Classe2 :** les paires de phrases ayant une similarité entre 0.5 et 0.80 partagent certains détails entre eux ;

- **Classe3 :** les paires qui diffèrent syntaxiquement et sémantiquement appartiennent à la troisième classe avec une similarité inférieure à 0.5. Elles peuvent être considérées comme un groupe de phrases partagent le même contexte d'étude.

A ce stade, nous avons terminé de collecter et traiter notre corpus selon l'approche proposée. Tous les résultats sont conservés dans des fichiers textes de tel sort que les fichiers contiennent les pairs de phrases et leurs similarités.

La prochaine étape entame l'évaluation de performance de notre technique de collection de données.

4. Évaluation et performance de l'approche proposée

L'étape de l'évaluation sert à mesurer la qualité et la fiabilité de données résultantes depuis l'application de l'approche de création de corpus parallèles monolingues de paraphrases en se basant sur les connaissances web.

En effet, pour des contraintes de temps et des matériels, nous avons effectué le test de notre approche sur un seul domaine spécialisé qui est le Cyber Criminalité, mais le résultat peut être généralisé sur tout autre domaine. Les caractéristiques de corpus de Cyber obtenu sont représentées dans le **tableau 5** suivant :

Nom du corpus	**Le nombre minimal de mots par phrase**	**Le nombre maximal de mots par phrase**	**Nombre de paires de Classe1**	**Nombre de paires de Classe2**	**Nombre de paires de Classe3**	**Total**
Cyber Criminalité	3 mots	50 mots	37844	140142	50253	228239

Tableau 5 : les statistiques de corpus de paraphrases 'Cyber'.

Le corpus présenté dans le « **tableau5** » a été assemblé en se basant sur notre approche en utilisant les méthodes exposées dans la partie traitant la conception et les algorithmes utilisés. Nous avons pu collecter 228239 couples de phrases depuis différents sites d'actualités sur Google et Wikipedia. Ces couples reparties sur trois classes (le graphe dans la **figue 13** suivante permet de visualiser les classes). La Classe1 sera utilisée comme dataset d'entrainement d'un générateur automatique de paraphrases.

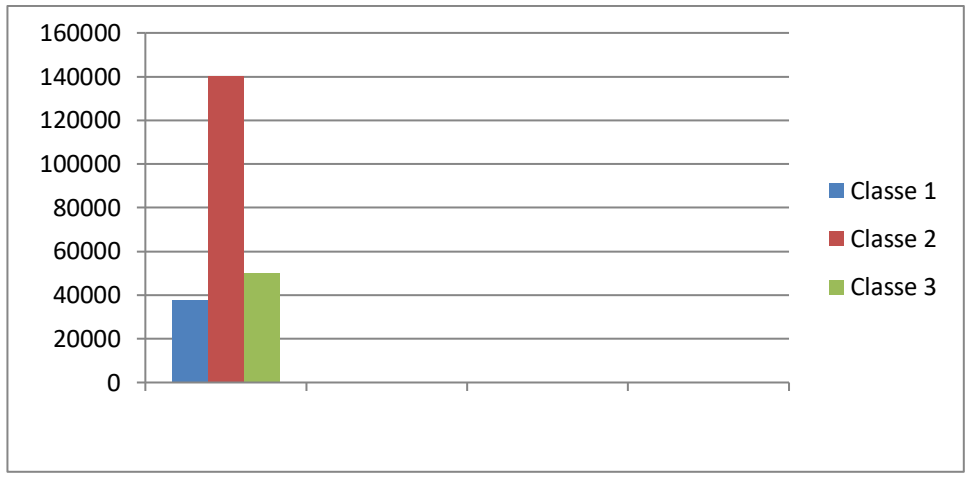


Figure 13 : représentation graphique de trois classes de corpus cyber.

Pour évaluer le travail réalisé nous avons effectué deux évaluations :

- Une évaluation qualitative manuelle intrinsèque sur le corpus construit,
- Une évaluation extrinsèque du corpus par rapport à la tâche de paraphrase.

Nous présentons dans la suite les deux évaluations :

4.1. Évaluation qualitative manuelle intrinsèque sur le corpus construit

Pour approuver les résultats, une évaluation humaine est toujours nécessaire. Un grand pourcentage d'applications TAL est évalué en demandant à des annotateurs humains ou à des experts du domaine pour mesurer la qualité des résultats.

Dans notre cas, Nous avons sélectionné un échantillon de 127 paires de phrases aléatoires (**Voir l'annexe A**) et nous l'avons affecté aux deux annotateurs qui sont des experts en langue arabe pour une évaluation humaine qualitative. Il a été demandé aux annotateurs de classifier les couples selon 5 classes :

Classe1 : Les deux phrases sont complètement semblables et donc paraphrases ;

Classe2 : Les deux phrases sont fondamentalement les mêmes (certains détails manquent) ;

Classe3 : Les deux phrases partagent quelques détails ;

Classe4 : Les deux phrases n'ont qu'un point commun qui est le sujet ;

Classe5 : Les deux phrases sont complètement différentes.

Les résultats sont représentés dans le graphe suivant :

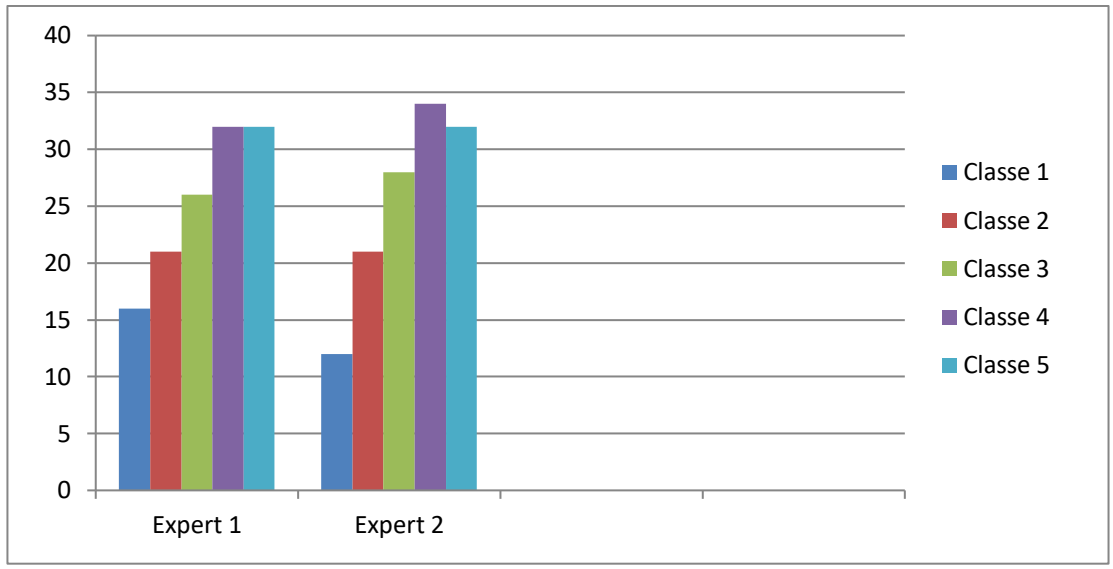


Figure 14 : représentation graphique des réponses des experts.

En calculant la corrélation entre réponses des experts (nous avons compté le nombre de fois ou les experts donnent la même réponse pour chaque couple de l'échantillon), nous trouvons un pourcentage de 80%, c'est une valeur assez importante par rapport à notre ensemble de données, cela signifie que les paires de phrases de notre corpus sont alignés d'une manière correcte. A partir de seuil de 0.8, les couples de phrases sont considérés des paraphrases (nous pouvons améliorer le processus de calcul de similarité avec une annotation manuelle plus poussée). Notamment, la sémantique de paraphrases entre les couples est assez préservée globalement, donc nous pouvons dire que notre approche a approuvé sa fiabilité dans la création des corpus de paraphrases.

4.2. Évaluation extrinsèque du corpus par rapport à la tâche de paraphrase

La génération automatique de paraphrases est un domaine de recherche en linguistique informatique qui étudie la possibilité d'attribuer la capacité à une machine de produire des paraphrases clair et similaire au sens de celle de la phrases sources. Cependant, le problème de manque de ressources constitue encore un défi à soulever pour la génération automatique et pour toute autre activité du TAL en général. Pour cette raison, nous avons proposé une approche de création de grands corpus de paraphrases depuis le web. Notre but n'est pas de créer de corpus seulement, mais aussi de l'utiliser pour entrainer des générateurs de paraphrases ou autre.

Dans cette partie nous allons illustrer l'architecture du système de génération automatique de paraphrases que nous cherchons à évaluer, et montrer l'emplacement de notre dataset dans

cette architecture globale.

Par la suite, nous passons à la partie expérimentale où, nous illustrons notre jeu d'essai utilisés dans notre test, ainsi les métriques d'évaluation quantitative et qualitative utilisées. Nous terminons par exposer et discuter les résultats obtenus.

4.2.1. Architecture globale du système de génération automatique

Le Système de génération automatique de paraphrases dont nous cherchons à évaluer les performances, est un outil proposé par [48], dans le cadre d'un projet de fin d'étude réalisé au cours de l'année 2019/2020, dédié à la langue Arabe. C'est un outil basé sur les réseaux de neurones approfondis. Il permet de générer des paraphrases cible à partir des phrases sources en utilisant un Encodeur/ Décodeur avec un mécanisme d'attention « EDAM ». Ce système est composé de deux parties principales, la première concerne le chargement et la mise en forme du dataset, et la seconde c'est le générateur de paraphrases. La sortie de ce système est un ensemble de phrases équivalentes sémantiquement à des phrases entrées. L'architecture globale de ce système est représentée dans la figure suivante :

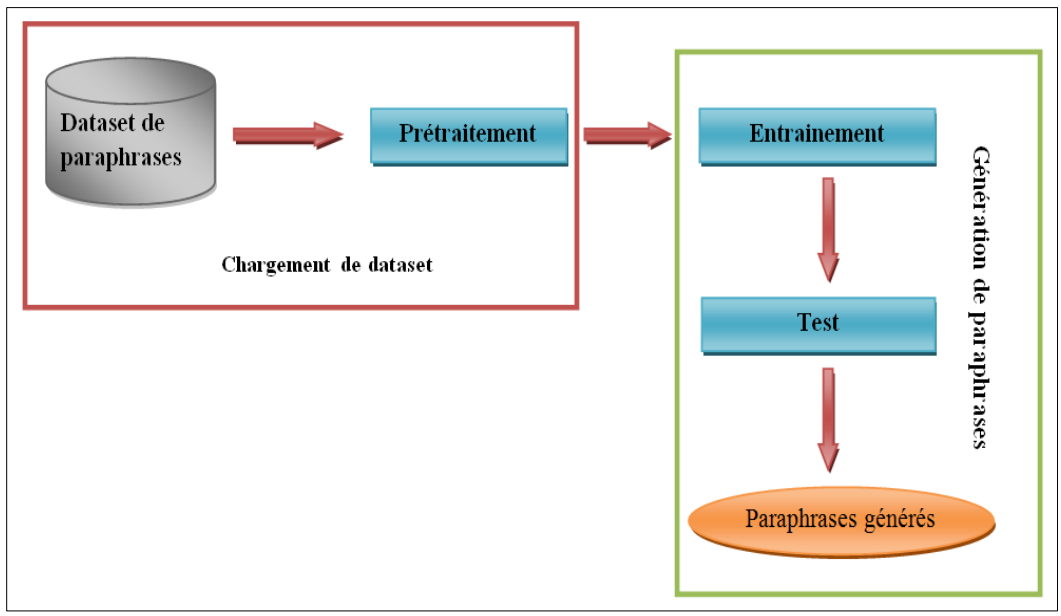


Figure 15 : architecture globale du système de génération de génération automatique de paraphrases [48].

A. La phase de chargement du dataset

Cette phase sert à alimenter le système par les données essentielles pour l'entraînement. Il comprend une étape consiste à transformer des données brutes, en format plus adapté et utilisable par le modèle, cette étape de « prétraitement » regroupe certaines tâches tels que la

normalisation, le formatage de lettres, la suppression des stops words…etc., expliquées en détaille dans les parties précédentes.

B. La phase de génération de paraphrases

Elle est divisée en deux étapes principales :

a. Entrainement : la partie entrainement sert à former un modèle de prédire des résultats à partir un ensemble de données, dans ce cas, il s'agit d'un apprentissage supervisé[19]. Lors de cette étape, un encodeur/décodeur (ED) avec un mécanisme d'attention 'EDAM' est utilisé, il s'agit d'un modèle de réseau neuronal qui permet de générer une sortie de séquence pour une entrée de longueur variable. Il est composé de deux éléments :

- **Encodeur** : est le générateur du vecteur contenant les informations relatives aux entrées, il sert à transformer les séquences d'entrée en des vecteurs tout en gardant le contexte de mots.
- **Décodeur** : le décodeur repose en entrée initialement sur le dernier état généré par l'encodeur qui contient les informations essentielles contenues dans chaque mot de la séquence d'entrée. A son tour, il analyse les vecteurs en entrée et cherche à prédire les mots correspondants en sortie.

L'intégration du mécanisme d'attention à l'ED sert à soulever le problème de taille fixe des vecteurs sortis de l'encodeur qui est incapable de retenir les informations contenant dans des séquences longues. Il permet d'améliorer la performance de modèle en mettant en valeur toutes les informations contenues dans les états cachés aux différents pas de temps au lieu de prendre celles contenues dans le dernier état caché (utilise tous les états cachés (contenant l'information de début d'encodage jusqu'à la fin) de l'encodeur pour générer un vecteur de contexte à chaque pas de temps).

Le principe de fonctionnement d'EDAM est illustré dans la figure 14 suivante :

[19]L'apprentissage supervisé consiste en l'entraînement d'une machine en utilisant des données labellisées. C'est-à-dire des données qui ont déjà été étiquetées avec le bon label (classe, valeur continue…).

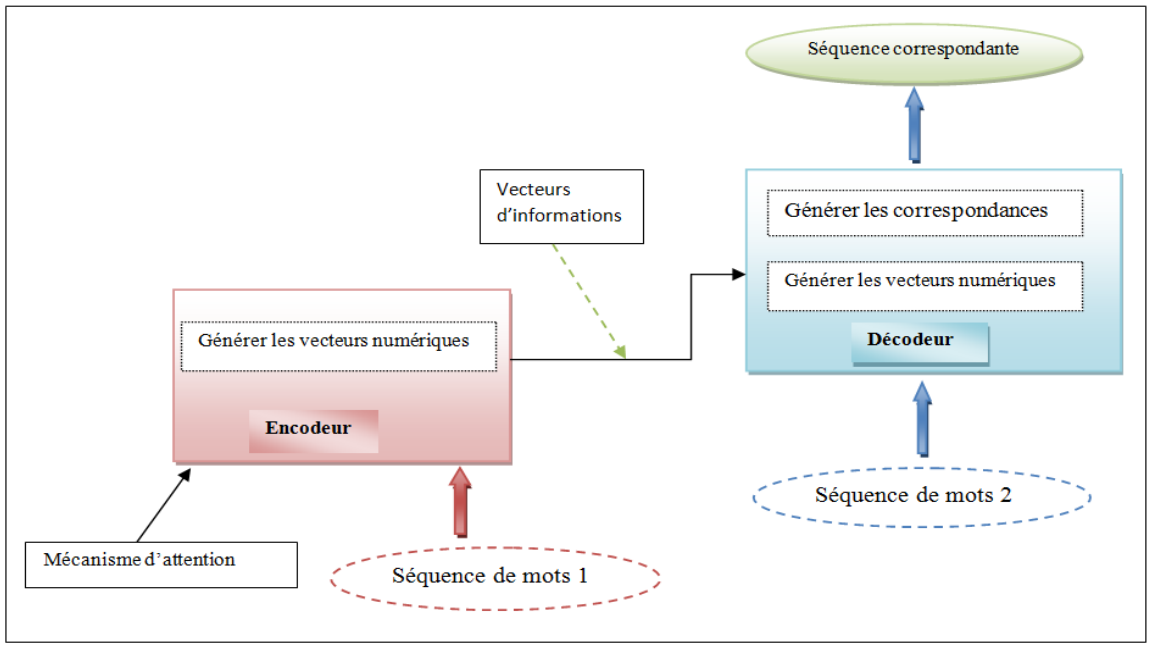


Figure 16 : le principe de fonctionnement d'EDAM [48].

b. Test : cette partie permet de mesurer la performance d'un modèle sur des données de test et permet donc de mesurer l'erreur du modèle final sur des données qu'il n'a jamais vues. C'est l'étape dans laquelle le modèle commence à tester sa base de connaissances acquises de processus de l'entrainement.

4.2.2. Expérimentation en utilisant notre dataset

Dans le cadre de notre projet, nous voulons entraîner le modèle EDAM à paraphraser automatiquement des phrases en utilisant l'ensemble de données que nous avons extrait au cours de ce travail. Le but de cette expérience est d'évaluer l'efficacité de notre jeu de données en apprentissage automatique pour générer automatiquement les paraphrases.

- ***Présentation de Dataset***

Le dataset utilisé dans cette phase est celui que nous avons créé au cours de ce travail. Nous avons sélectionné l'ensemble de données dont l'évaluation automatique a montré un degré élevé de similitude entre les couples de phrases. Nous allons consacrer 80% de jeu de données pour l'entrainement et 20% pour le test comme illustré dans le **Tableau 6**.

Dataset	Entrainement (80%)	Test (20%)	Total (100%)
Cyber	30275	7569	37844

Tableau 6 : détails du dataset.

- ***Génération de paraphrases***

Nous avons fait l'entrainement de générateur EDAM avec 30275 paires de phrases (phrases sources / paraphrases). Pour garantir la qualité d'apprentissage, nous avons lancé l'entrainement plusieurs fois pour vérifier que le pourcentage de perte d'information« loss » était complètement réduit. Ici, il est très important d'avoir un grand jeu de données, ce qui permet au modèle de bien apprendre et de donner des résultats pertinents. Cette étape prend beaucoup de temps et consomme beaucoup de ressources de la machine, pour cela nous l'avons réalisé sur Google Colab[20], mais nous avons rencontré souvent le problème de déconnexion de l'outil lors de l'exécution ce qui nous obligeons de refaire l'opération à chaque fois.

Le test est fait par le 20% qui reste. Le système prend en entrée les 7569 phrases sources et prédit leurs correspondances selon sa base de connaissances qu'il a formée plutôt dans la phase d'apprentissage.

A ce stade, nous avons trois genres de phrases :

Une phrase source : la phrase que nous avons utilisée pour la génération ;

Une paraphrase référence : qui est la paraphrase initiale dans notre jeu de données ;

Une paraphrase hypothèse : qui est la paraphrase générée par le modèle EDAM.

Un exemple de paraphrases générées par le modèle EDAM est présenté ci-après :

Phrase source : الجريمة الإلكترونية ذات بعد دولي فهي قد تتجاوز الحدود الجغرافية

Paraphrase référence : جريمة معلومات عابرة للحدود

Paraphrase hypothèse : الجريمة الإلكترونية عابره للحدود

4.2.3. Evaluation des résultats

Le processus d'évaluation est une tâche très importante car il nous permet de connaître la qualité des paraphrases obtenues et nous donne un aperçu de la performance de nos approches et des méthodes utilisées. Pour cette fin, nous avons proposé de mesurer quantitativement et qualitativement la qualité de paraphrases générées automatiquement par

[20]Colaboratory est un environnement de notebook Jupyter gratuit qui ne nécessite aucune configuration, il permet d'exécuter des codes python sur le navigateur.

le modèle EDAM entrainé par notre dataset.

4.2.3.1. L'évaluation quantitative

L'évaluation automatique consiste à calculer le taux de similarité entre la phrase générée (paraphrase hypothèse) et celle du dataset (paraphrase référence).Pour ce faire, plusieurs métriques peuvent être utilisées. Nous avons retenue la plus utilisée à savoir BLEU et GLEU.

- **BLEU** (bilingue évaluation understudy) [49] : BLEU compte les N-grammes correspondants dans la traduction générée en N-grammes dans le texte d'or ou de référence. Ici, l'unigramme est un jeton et le bi-gramme est une paire de mots.

La sortie de BLEU est toujours un nombre compris entre 0 et 1. Cette valeur indique à quel point le texte candidat est similaire aux textes de référence, avec des valeurs plus proches de 1 représentant des textes plus similaires.

- **GLEU** (Google-BLEU) : est une variante de la métrique BLEU, utilisée spécialement pour mesurer taux de corrections d'erreurs grammaticales des n-grammes générés avec l'ensemble des phrases références [50].

Le résultat d'application des deux métriques est représenté dans le tableau suivant :

Métrique	BLEU	GLEU
score	36%	18%

Tableau 7 : les scores de BLEU et GLEU.

L'interprétation des résultats des scores BLEU est reprise de l'évaluation des modèles proposés par Google [51] **tableau 8** :

Score Bleu (%)	Interprétation
<10	Résultat presque inutile
10 à 19	L'idée générale est difficilement compréhensible
20 à 29	L'idée générale apparait clairement, mais le texte comporte de nombreuses erreurs grammaticales
30 à 40	Résultats compréhensibles et correctes.
40 à 50	Résultat de haute qualité
50 à 60	Résultat de très haute qualité, adéquat et fluide
> 60	Qualité souvent meilleure que celle donnée par l'humain.

Tableau 8 : interprétation des scores BLEU.

Selon ces interprétations nous pouvons dire que le modèle EDAM a donné un score BLEU assez comparable (36%) et les résultats sont compréhensibles et correctes.

Nous pouvons juger aussi la qualité des résultats par les métriques d'évaluations automatiques fournissent des mesures utiles sur la qualité du langage (grammaire et forme). Cependant, elles ne fournissent aucune mesure sur la pertinence du contenu obtenu (la sémantique) [52].

Pour quantifier ces aspects qui ne sont pas abordés par les métriques d'évaluation automatiques, l'évaluation humaine devient nécessaire pour notre problème à savoir la génération de phrases ayant le même sens qu'une phrase originale.

4.2.3.2. Evaluation Qualitative

Les évaluations basées sur des jugements humains sont nécessaires et complémentaires aux évaluations des métriques automatiques car cette dernière ne suffit pas pour évaluer les paraphrases dans une perspective fine, en termes de deux aspects :

- La « **Relevance** » (Pertinence en sens) : exprime la pertinence de la paraphrase générée avec la phrase d'entrée. Ici il est question de noter à quel point la phrase générée préserve le même sens que la phrase originale.
- La « **Readability** » (lisibilité en forme) : la lisibilité de la paraphrase générée en termes de forme, de grammaire sans considérer le sens de la phrase générée.

Cette évaluation est appliquée sur un échantillon de 100 couples (phrases originales, paraphrases générées par le modèle) sélectionnées aléatoirement d'un ensemble de couples.

Pour chacun des deux aspects, un annotateur attribue un score compris entre 1 et 5, où 1 est le pire et 5 est le meilleur.

Nous avons fait une évaluation qualitative pour les mêmes couples par trois experts humains volontaires, maitrisant la langue et nous garderons la moyenne des notes obtenues pour chaque couple.

Dataset	**Relevance/5**	**Readability/5**
Cyber	3	4,01

Tableau 9 : résultats d'évaluation humaine qualitative.

D'après le **tableau 9**, nous constatons les points suivants :

Les annotations des experts arabes sont assez bonnes (60% du sens est préservé et la lisibilité des paraphrases généré est à 80.2%).

Les résultats des annotations manuelles des experts confirment encore une fois l'importance d'accompagner aux évaluations automatiques des annotations humaines pour garder les deux aspects de Readability et la Relevance.

Les résultats obtenus nous a permet de mettre en valeur les points suivants :

La taille du dataset est un facteur important qui influence directement sur la qualité des paraphrases générées automatiquement. Plus la taille est grande, meilleur résultat il est.

Les annotations manuelles nous permettent de constater que la qualité des paraphrases générées et assez bon en terme de sens et lisibilité.

5. Difficultés rencontrées :

Durant notre travail, nous avons rencontrés plusieurs problèmes avant d'arriver à réaliser ce travail, nous citons dans cette partie ces différentes difficultés :

- A noter que nos corpus sont générés à partir du web ou de Wikipédia, donc, une connexion internet fiable est nécessaire ce qui n'a pas toujours été le cas.
- Certains sites retournent des liens sans contenus textuels, spécialement les sites Google, comme le cas des liens de vidéos. Pour éviter le maximum de ce genre de problème, il est préférable de fournir des mots clés plus précis.
- Les traitements de corpus nécessitent des machines plus puissantes que nous n'avons pas. La tâche s'avère dur sur une machine de mémoire de 4Gb. Durant cette phase, notre PC est tombé en panne plusieurs fois.
- L'étape de calcul de similarité prend un certain temps surtout si nous utilisons notre propre machine local. Le calcul, par exemple, de similarité de deux fichiers contenant 10 phrases chacun peut prendre jusqu'à 25 min comme le montre la **figure 17** suivante. Pour résoudre ce problème, nous avons exploité l'outil Google Colab[21] afin d'avoir plus de mémoire pour aboutir aux résultats, nous avons pu avancer mais ça reste quand même assez limité.

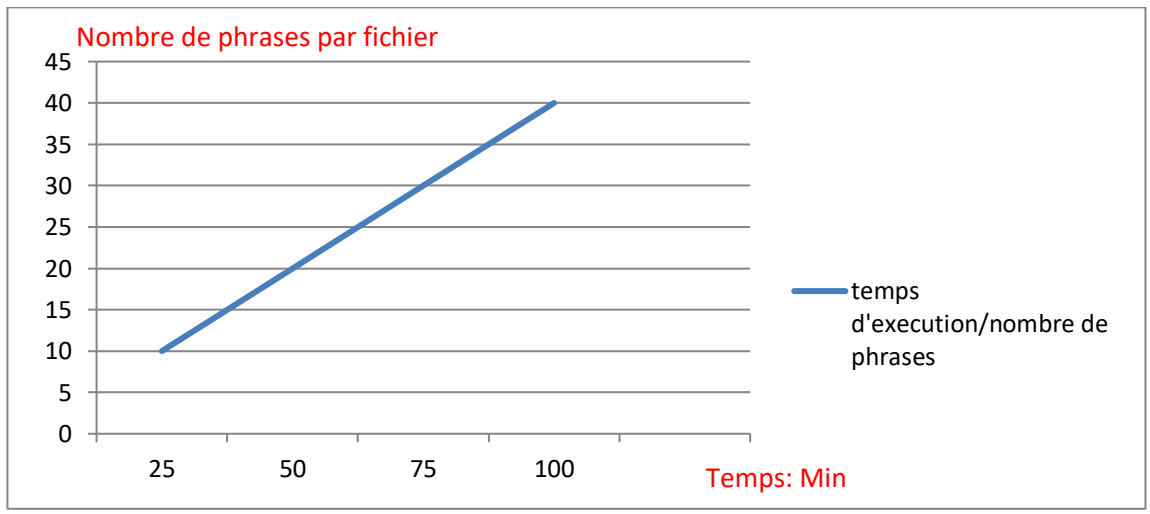


Figure 17 : le temps de calcul de similarité en fonction de taille de fichiers de phrases.

[21] Colaboratory est un environnement de notebook Jupyter gratuit qui ne nécessite aucune configuration, il permet d'exécuter des codes python sur le navigateur.

Conclusion

Dans cette partie, nous avons présenté notre approche de création de corpus de paraphrases depuis le web. Nous avons fini de structurer notre démarche et de l'implémenté réellement. L'évaluation qualitative par des experts de notre corpus nous a donnés des résultats satisfaisants et montrés la pertinence de notre méthodologie.

Dans le chapitre prochain, nous allons présenter une application développée pour consolider l'approche étudiée.

Chapitre 3 : Consolidation des outils développés

1. Introduction

Il est clair que le traitement automatique de langage naturel nécessite une très grande masse de données pour assurer la fiabilité de ses résultats. Notamment, la tâche de collection et traitement de données s'avèrent très difficile lors de l'application manuellement. Une façon de pallier ce problème est de développer des outils consistent à automatiser ce processus, et c'est l'un de nos objectifs.

Dans cette partie, nous allons présenter '**AP-Corpora'**, notre consolidation d'outils pour l'acquisition automatique de corpus de paraphrases.

2. Environnement de travail

L'environnement de travail est constitué par deux parties nommées environnement matériel et environnement logiciel.

2.1. Environnement matériel

Le développement de l'environnement matériel est caractérisé par :

1. Système d'exploitation : Linux Ubuntu 20.04 ;
2. CPU : Intel, Core i3-4005U, 1.70GHz × 4 ;
3. Mémoire : 3.8 GB ;
4. Capacité : 500.1 GB.

2.2. Environnement logiciel

L'environnement logiciel consiste les composants suivants :

•Python 3.8.10 ;

•Pycharm Commuity Edition 3.6.

3. Présentation de l'application et fonctionnalités :

'AP-Corpora' est un outil simple et basique de collection et de traitement de données dans lequel nous avons intégré toutes les tâches nécessaires définies dans les parties précédentes permettant la création de corpus de paraphrases. Cet outil est développé en utilisant python avec la bibliothèque Tkinter[22].

[22]Tkinter est l'interface Python standard de la boîte à outils Tcl/Tk GUI, disponible sur https://docs.python.org/3/library/tkinter.html.

Le présent système comprend les fonctionnalités suivantes :

a. Gérer les corpus

Dans cette interface, l'utilisateur à la possibilité d'ajouter, modifier, supprimer et mettre à jour les corpus. Depuis cet écran il aura la liste de ses corpus, la capture suivante montre ces fonctionnalités.

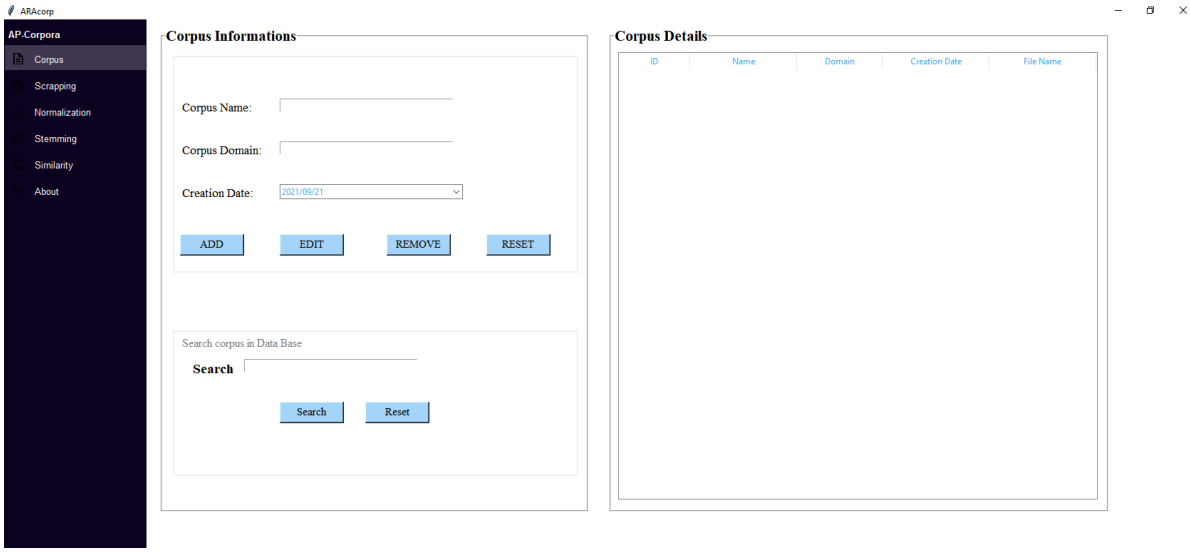


Figure 18 : interface de gestion de corpus.

b. Scraper le Web

Cette fonction permet à l'utilisateur d'effectuer une requête pour interroger le moteur de recherche « voir la figure suivante ».

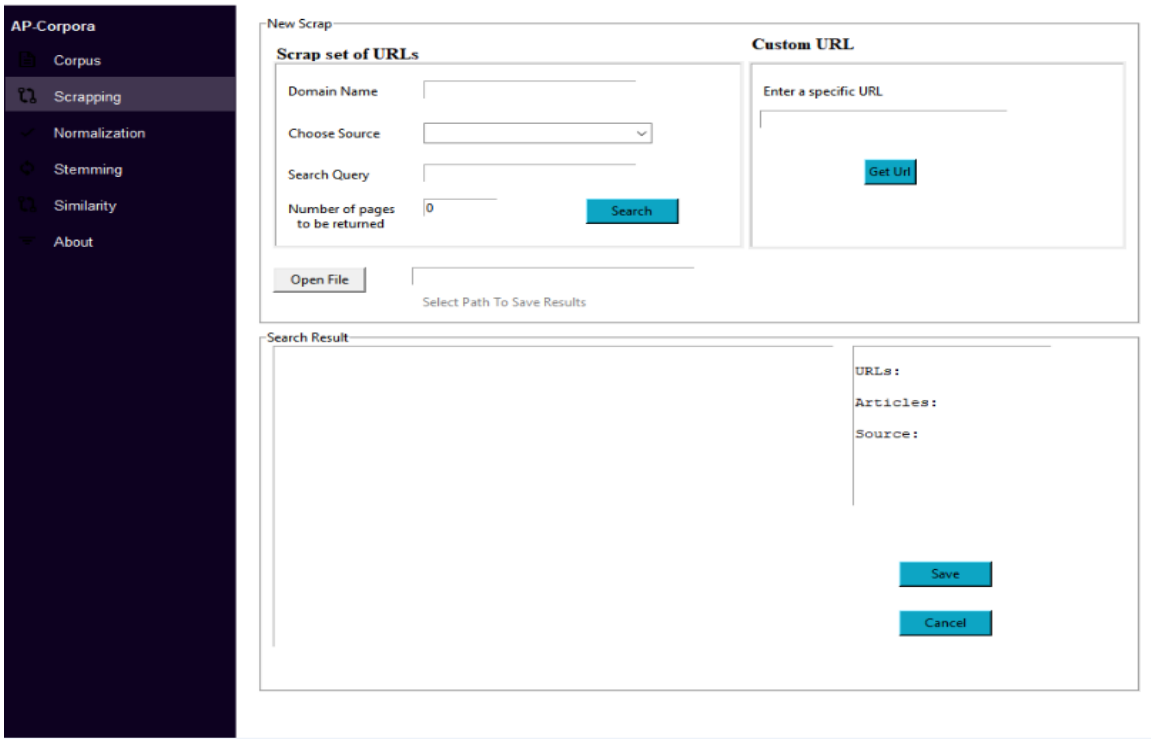


Figure 19 : interface de scraping.

Dans cette interface, l'utilisateur va avoir deux choix à savoir :

Scraper le web en effectuant une requête avec mots clés, dans ce cas-là, il est invité à spécifier le domaine de recherche, la source de données (Google ou Wikipedia), le nombre de pages à retourne ainsi la requête désirée.

Scraper une url spécifique : il possible de donner une url personnalisée et le système va le chercher.

Dans les deux cas, le système va transmettre une requête au moteur de recherche et il va retourner les résultats sous forme de textes, ces résultats serons enregistrer dans des fichiers textes l'un indépendamment de l'autre pour pouvoir les récupérés par la suite.

c. Normalisation de données

Notre système offert la possibilité de normaliser les données en utilisant plusieurs critères présentés dans l'interface suivante « figure ». Notamment, l'utilisateur va sélectionner un fichier texte et choisir le type de prétraitement souhaité à savoir : la suppression des signes diacritiques, élimination des Stop Words, suppression des mots non arabe et symboles ainsi la normalisation des lettres arabe.

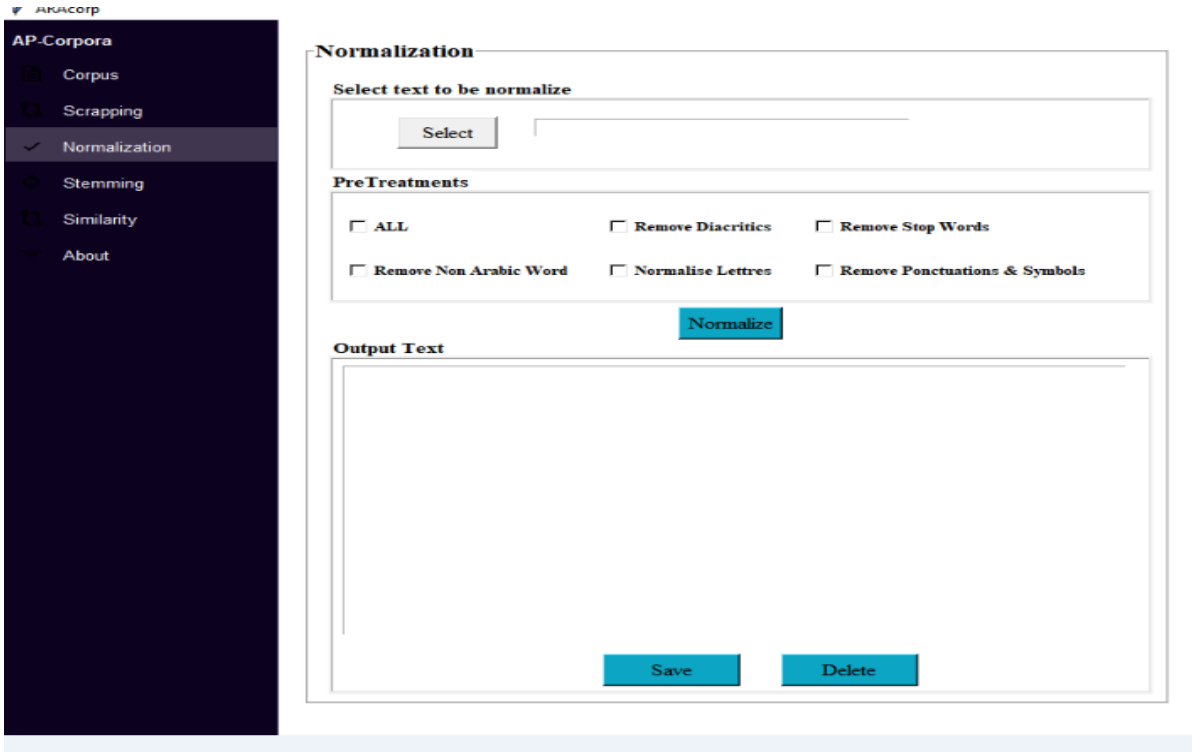


Figure 20 : interface de normalisation.

d. Stemmer un corpus

Dans cette partie ; nous donnons à l'utilisateur la possibilité de stemmer son corpus en utilisant les deux méthodes ISRI et Tashaphyne comme le montre l'interface suivante.

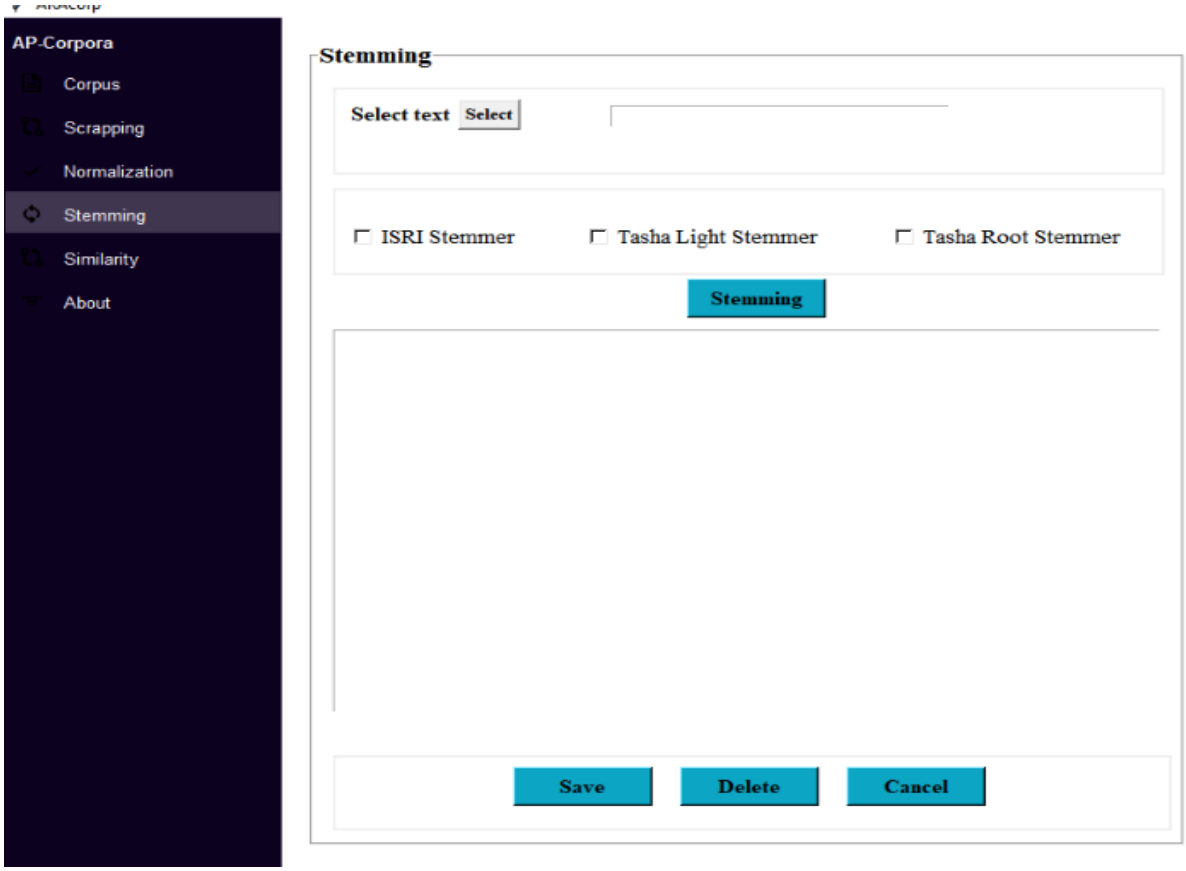


Figure 21 : interface de stemming.

e. Calculer les similarités

Avec cet outil, l'utilisateur va avoir plusieurs fonctionnalités tels que :

Calculer la similarité entre deux fichiers texte contenant des phrases normalisées.

Calculer la similarité entre deux phrases.

Dans les deux cas, il est invité de sélectionner les mesures de similarité qui lui convient parmi la liste présentée au-dessus de l'interface (similarité sémantique et similarité syntaxique). Ainsi, il va avoir la possibilité de filtrer directement les paires de phrases obtenues en sélectionnant l'un des méthodes de filtrage (pondération, moyenne harmonique). Ces différentes fonctionnalités sont représentées dans la capture d'écran suivante :

Figure 22 : interface de calcul de similarité.

Conclusion générale & Perspectives

Conclusion

Nous nous sommes intéressés dans ce travail à proposer une méthodologie de création de grands corpus parallèles monolingues de paraphrases**.** Nous avons développé une approche purement destinée au domaine du TAL, permet d'exploiter une ressource de données non encore utilisée pour l'arabe.

Notre méthode était concentrée sur l'exploitation du web comme une ressource de données pour créer des grands corpus de paraphrases.

Nous avons commencé par un état de l'art dans lequel nous avons présenté le contexte d'étude, les concepts fondamentaux liées à notre recherche (TAL arabe, la notion de paraphrase, les corpus de paraphrases, …).Ainsi, nous avons fait une étude sur les travaux connexes à notre travail. Nous avons conclu que la majorité de ces travaux ne traitent pas la langue Arabe. Ce qui nous donne une motivation de développer de plus cette technique car il nous semble que cette approche peut être prometteuse dans les différentes domaines du TAL.

A partir l'étude de travaux existants, nous avons commencé à concevoir et réaliser notre approche basée sur les connaissances Web (Web-knowledge). Nous avons procédé à créer un grand corpus monolingue de paraphrase arabe dédié pour entrainer un modèle de génération automatique de paraphrases. Notre approche donne l'importance aux objectifs suivants :

- Faciliter le processus de création de corpus de paraphrases et en donner une méthodologie générique ;
- Consolider les travaux de recherche de TAL arabe ;
- Offrir un grand ensemble de données pour entrainer un générateur de paraphrases.

Le corpus que nous avons créé est évalué par des experts en Arabe. Ces évaluations ont mis en valeur notre approche. Une amélioration de méthodes de calcul de similarité est importante pour assurer un meilleur filtrage des couples.

L'intégration de notre Dataset dans le processus de génération automatique de paraphrases a donné des résultats satisfaisants qui nous ont encouragés pour améliorer de plus notre approche et la généraliser pour plusieurs domaines afin d'augmenter plus la taille du corpus.

Signalons par ailleurs vu les difficultés rencontrées en terme de ressources matérielles lors de la réalisation du travail que des services de laboratoire de recherche bien équipés deviennent indispensables pour mener ce genre de recherches (stations de travail, serveurs dédiés, ….).

Perspective

Les corpus de paraphrases sont importants dans plusieurs domaines du TAL, pour cela nous mettons comme perspective d'aller plus profond dans ce travail et enrichir de plus les collections de données avec une variété de domaines. Ainsi, optimiser le temps de calcul de similarités pour améliorer notre approche.

Références bibliographiques

[1] Bouamor, H., Max, A., & Vilnat, A. (2010, August). Comparison of paraphrase acquisition techniques on sentential paraphrases. In *International Conference on Natural Language Processing* (pp. 67-78). Springer, Berlin, Heidelberg.

[2] Fuchs, C. (1987). L'ambiguïté et la paraphrase. Opérations linguistiques /http://languesweb.blogspot.com/2014/12/lambiguite-et-la-paraphrase-almuth.html

[3] Benhamida.,A. , « Vers une plateforme de gestion des corpus et d'analyse de texte en langue arabe. ». Mémoire master USDB 1. 2019/2020.

[4] Barzilay, R., & McKeown, K. (2001, July). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics* (pp. 50-57).

[5] Shinyama, Y., Sekine, S., Sudo, K., & Grishman, R. (2002, March). Automatic paraphrase acquisition from news articles. In *Proceedings of HLT* (Vol. 2, p. 1). San Diego, US.

[6] Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*.

[7] Lin, D., & Pantel, P. (2001, August). Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 323-328).

[8] Bouamor, H., Max, A., & Vilnat, A. (2012). Étude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques. *TAL*, *53*(1), 11-37.

[9] Fuchs, C. (1980). Quelques réflexions sur la paraphrase dans les théories du langage. *L'information grammaticale*, *6*(1), 37-44.

[10] Anchiêta, R. T., Sousa, R. F. D., & Pardo, T. A. (2020). Modeling the Paraphrase Detection Task over a Heterogeneous Graph Network with Data Augmentation. *Information*, *11*(9), 422.

[11] Pottier, B. (1989). La paraphrase textuelle dans ses fondements théoriques. *Cahiers d'Études Hispaniques Médiévales*, *14*(1), 37-45.

[12] Bhagat, R., & Hovy, E. (2013). What is a paraphrase?. *Computational Linguistics*, *39*(3), 463-472.

[13] Grabar, N., & Eshkol, I. (2015, June). ... des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de

corpus oraux. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs* (pp. 303-316).

[14] Sinclair, J. (1996). Preliminary recommendations on corpus typology. *EAGLES Document eag-tcwg-ctyp/p*.

[15] Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge University Press.

[16] McEnery T. et Gabrielatos C.(2006) ; English corpus linguistics. The Hand- book of English Linguistics, pages 33–71.

[17] Miftah, N., Allah, F. A., & Taghbalout, I. (2016). Corpus multilingues pour les langues peu dotees. In *proceedings of the 7th TICAM conference*.

[18] Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, *7*(2), 223-243.

[19] Sinclair, J. (1995). Corpus typology: A framework for classification. *Stockholm studies in English*, *85*, 17-33.

[20] Bannard, C., & Callison-Burch, C. (2005, June). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 597-604).

[21] Koehn, P. (2005, September). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (Vol. 5, pp. 79-86).

[22] Bernhard, D., & Gurevych, I. (2008, June). Answering learners' questions by retrieving question paraphrases from social Q&A sites. In *Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 44-52).

[23] Deléger, L., & Zweigenbaum, P. (2009, August). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)* (pp. 2-10).

[24] Callison-Burch, C., Koehn, P., & Osborne, M. (2006, June). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp. 17-24).

[25] Al-Raisi, F., Bourai, A., & Lin, W. (2018). Neural Symbolic Arabic Paraphrasing with Automatic Evaluation. *Computer Science & Information Technology*, *1*.

[26] Do, T. N. D. (2011). *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée* (Doctoral dissertation, Université de Grenoble; Université de Hanoi--Vietnam).

[27] Kajiwara, T., & Komachi, M. (2016, December). Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1147-1158).

[28] Klerke, S., & Søgaard, A. (2012, May). DSim, a Danish Parallel Corpus for Text Simplification. In *LREC* (pp. 4015-4018).

[29] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

[30] Bannard, C., & Callison-Burch, C. (2005, June). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 597-604).

[31] Ma, X., & Cieri, C. (2006, May). Corpus Support for Machine Translation at LDC. In *LREC* (pp. 859-864).

[32] Bouamor, H. (2010, July). Construction d'un corpus de paraphrases d'énoncés par traduction multiple multilingue. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues* (pp. 44-53).

[33] Al-Raisi, F., Lin, W., & Bourai, A. (2018). A monolingual parallel corpus of arabic. *Procedia computer science*, *142*, 334-338.

[34] Sakre, M. M., Kouta, M. M., & Allam, A. M. (2009). automated construction of Arabic-English parallel corpus. *Journal of the Advances in Computer Science*, *3*.

[35] Wubben, S., Van den Bosch, A. P. J., & Krahmer, E. J. (2014). Creating and using large monolingual parallel corpora for sentential paraphrase generation.

[36] Dolan, W., Quirk, C., Brockett, C., & Dolan, B. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.

[37] Madnani, N., & Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, *36*(3), 341-387.

[38] Al-Kabi, M. N., Kazakzeh, S. A., Ata, B. M. A., Al-Rababah, S. A., & Alsmadi, I. M. (2015). A novel root based Arabic stemmer. *Journal of King Saud University-Computer and Information Sciences*, *27*(2), 94-103.

[39] Adduru, V., Hasan, S. A., Liu, J., Ling, Y., Datla, V. V., Qadir, A., & Farri, O. (2018, January). Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In *KHD@ IJCAI*.

[40] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed

representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

[41] Levenshtein, V., (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Sov. Phys. Dokl. 10, 707–710.

[42] Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414-420.

[43] JACCARD P. (1912). The Distribution of the Flora in the Alpine Zone.1. New Phytologist 11 (2): 37– 50. https://doi.org/10.1111/j.1469 8137.1912.tb05611.x.

[44] DICE L. R. (1945). "Measures of the Amount of Ecologic Association Between Species." Ecology 26 (3): 297–302. https://doi.org/10.2307/1932409.

[45] Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC medical informatics and decision making*, *18*(1), 1-13.

[46] Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, *112*, 340-349.

[47] Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., & Atyia, A. (2015, April). Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 430-443). Springer, Cham.

[48] Lamari, S., Hamel, O., (2020). " Les Réseaux De Neurones Pour La Génération Automatique De Paraphrases", Mémoire de master, Université Saad Dahlab blida 1.

[49] Wołk, K., & Marasek, K. (2015). Enhanced bilingual evaluation understudy. *arXiv preprint arXiv:1509.09088*.

[50] Palivela, H. (2021). Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, *1*(2), 100025.

[51] Évaluer des modèles | Documentation d'AutoML Translation." [Online]. Available: https://cloud.google.com/translate/automl/docs/evaluate?hl=fr&fbclid=IwAR3f_J5uf2 J5uf2uW_x-sKlboncA-wbXVMJ4FsPFxnT-tjEljPALWBM9uSp8ZjD0. [Accessed: 22-Sept-2021].

[52] Babych, B. (2014). Automated MT evaluation metrics and their limitations. *Tradumàtica*, (12), 0464-47

Annexe –A-

I. Extrait de l’échantillon de paires de phrases envoyées aux annotateurs pour l’évaluation du dataset

الجملة 1	الجملة 2	متشابهة تماما	متشابهة بشكل أساسي (بعض التفاصيل مفقودة)	مشاركة بعض التفاصيل (الكثيرمن التفاصيل مفقودة)	مشاركة الموضوع فقط	مختلفة تماما	ملاحظة اخرى
أعلن وزير الخارجية الأميركي أنتوني بلينكن، اليوم الاثنين، أن قراصنة متعاونين مع وزارة أمن الدولة الصينية، هم الذين نفذوا عملية التجسس الإلكترونية الضخمة التي استهدفت في وقت سابق هذا العام آلاف أجهزة الحاسوب الكمبيوتر التابعة لمؤسسات حكومية أميركية، مستغلين ثغرة في أنظمة مايكروسوفت	عودة ابتزاز الفدية وقالت شركة كوب سويدن التي تمثل نحو 20 من القطاع في البلاد ويناهز حجم مبيعاتها 1 5 مليار يورو في بيان تعرض أحد مقاولينا لهجوم الكتروني، ما أدى إلى توقف عمليات الدفع لدينا عن العمل						
وأضاف أن واشنطن ستتخذ إجراءات ضد المتورطين في الأنشطة السيبرانية الخبيثة، بسبب سلوكهم غير المسؤول في الفضاء الإلكتروني	ونسبت الشرطة الفيدرالية الأمريكية هذه الهجمات إلى قراصنة على الأراضي الروسية، وتنفي موسكو التغطية على أنشطتهم أو حتى الارتباط بهم						
وفي بداية لمرحلة جديدة من التوتر مع الصين، انضمت أميركا إلى حلف شمال الأطلسي ناتو والاتحاد الأوروبي وبريطانيا وأستراليا واليابان ونيوزيلندا وكندا لتوجيه الاتهامات إلى بكين بشأن الهجوم السيبراني الكبير الذي اكتشف في مارس آذار الماضي، وذلك وفق وثيقة نشرها البيت الأبيض صباح اليوم	عودة ابتزاز الفدية وقالت شركة كوب سويدن التي تمثل نحو 20 من القطاع في البلاد ويناهز حجم مبيعاتها 1 5 مليار يورو في بيان تعرض أحد مقاولينا لهجوم الكتروني، ما أدى إلى توقف عمليات الدفع لدينا عن العمل						

وتظهر إحصائيات وحدة الجرائم الإلكترونية التابعة لمديرية البحث الجنائي في مديرية الأمن العام ازدياداً مضطرداً في أعداد الجرائم الإلكترونية التي تتعامل معها الوحدة في كل عام	رصدت الشرطة الجنائية الدولية أنتربول نحو 907 آلاف هجوم سيبراني غير مرغوب فيه مرتبط بفيروس كورونا تعرضت له العديد من الدول، في مسح شارك فيه المغرب من ضمن مجموعة من الدول						
تعرّف الجرائم الإلكترونية بأنها الأنشطة الإلكترونية التي ترتكب عمدا بدافع إجرامي لإلحاق ضرر مادي أو معنوي بشخص ما أو جهة ما، بشكل مباشر أو غير مباشر	وأحيانا توصف الأنشطة التي تتعلق بالدول وتُستهدف فيها دولة أخرى واحدة على الأقل بأنها تقع في إطار الحرب الإلكترونية ، والنظام القانوني الدولي يحاول تحميل الفاعلين المسؤولية عن أفعالهم في مثل هذا النوع من الجرائم من خلال المحكمة الجنائية الدولية						
ولا تقتصر الجرائم الإلكترونية على أفراد أو مجموعات، وإنما قد تمتد إلى مستوى الدول فتهدد أمنها القومي وسلامتها المالية، ويشمل ذلك التجسس الإلكتروني وأبرز أمثلته ما كشفته تسريبات الأميركي إدوارد سنودن بشأن مخططات الإدارة الأميركية للتجسس على اتصالات الأفراد والدول الأخرى ، والسرقة المالية، وغيرها من الجرائم العابرة للحدود	وتشمل الجرائم الإلكترونية أي فعل إجرامي يتم من خلال الحواسيب أو الشبكات كعمليات الاختراق والقرصنة، كما تضم أيضا أشكال الجرائم التقليدية التي يتم تنفيذها عبر الإنترنت						

II. Extrait de l’échantillon de paraphrases générées par EDAM envoyé aux annotateurs pour évaluer la génération des paraphrases

	Source phrases (العبارات الأصلية المدخلة)	Generated paraphrases (العبارات المصاغة الناتجة)	Relevance (الملاءمة في المعنى)	Readability (المقروئية)
1	في التسعينات حيث شهدت هذه المرحله تناميا هائلا في حقل الجرائم الالكترونيه نظرا لانتشار الانترنت في هذه الفتره مما سهل من عمليات دخول الانظمه واقتحام شبكه المعلومات مثلا تعطيل نظام تقني نشر الفيروسات الخ	في السنوات الاخيره ازداد حجم وقوه الهجمات السيبرانيه بشكل سريع كما لاحظ المنتدى الاقتصادي العالمي مقابل 450 مليار دولار عام 2016		
2	نشر واستخدام برامج الحاسب الالي بما يشكل انتهاك لقوانين حقوق الملكيه و الاسرار التجاريه	يمكن استخدام نشر المعلومات غير الصحيحه		
3	جريمه عابره للحدود لا تعترف بعنصر المكان والزمان فهي تتميز بالتباعد الجغرافي واختلاف التوقيتات بين الجاني والمجني عليه فالسهوله في حركه المعلومات عبر انظمه التقنيه الحديثه جعل بالامكان ارتكابها عن طريق حاسوب موجود في دوله معينه بينما يتحقق الفعل الاجرامي في دوله اخرى	و يمكن تعريف الجريمه الالكترونيه على انها اي مخالفه ترتكب ضد افراد او جماعات بدافع جرمي و نيه الاساءه لسمعه الضحيه او لجسدها او عقليتها سواء كان ذلك بطريقه مباشره او غير مباشره و ان يتم ذلك باستخدام وسايل الاتصالات الحديثه مثل الانترنت غرف الدردشه البريد الالكتروني او المجموعات		
4	حسب منظمه التعاون الاقتصادي للجريمه المرتكبه عبر الانترنت هي كل سلوك غير مشروع او غير اخلاقي او غير مصرح به يتعلق بالمعالجه الاليه للبيانات ونقلها	في السنوات الاخيره ازداد حجم وقوه الهجمات السيبرانيه بشكل سريع كما لاحظ المنتدى الاقتصادي العالمي في تقريره لعام 2018 القدرات السيبرانيه الهجوميه تتطور بسرعه اكبر من قدرتنا على التعامل مع الحوادث العدابيه		
5	دوافع ذهنيه او نمطيه غالبا ما يكون الدافع لدى مرتكب الجرائم عبر الانترنت هو الرغبه في اثبات الذات وتحقيق الانتصار على تقنيه الانظمه المعلوماتيه دون ان يكون لهم نوايا اثمه	الامن السيبراني يطلق عليه ايضا امن المعلومات و امن الحاسوب وهو فرع من فروع التكنولوجيا يعنى بحمايه الانظمه والممتلكات والشبكات والبرامج من الهجمات الرقميه التي تهدف عاده للوصول الى المعلومات الحساسه او تغييرها او اتلافها او ابتزاز المستخدمين للحصول على الاموال او تعطيل العمليات التجاريه		
6	ان الحفاظ على امن المعلومات لا يتعلق بالاجراءات التقنيه فحسب وانما يجب النظر الى امن المعلومات من منظور اوسع لذلك يمكن القول ان اجراءات امن المعلومات يمكن ان تكون	تتعدد تعريفات امن المعلومات وتتنوع حسب زاويه الرويه فنحن اذا نظرنا من زاويه اكاديميه سنجد انه العلم الذي يبحث في نظريات واستراتيجيات توفير الحمايه للمعلومات من المخاطر التي تهددها ومن انشطه الاعتداء عليها		
7	ان امن المعلومات ليس مرتبطًا فقط بحمايه المعلومات من الوصول غير المرخص وانما يمتد ايضًا لمنع اي استخدام او كشف او اتلاف او تعديل او تفتيش او نسخ غير مصرح به للمعلومات	البيانات المعلومات او الاوامر او الرسايل او الاصوات او الصور التي تعد او التي سبق اعدادها لاستخدامها في الحاسب الالي وكل ما يمكن تخزينه ومعالجته ونقله وانشاوه بوساطه الحاسب الالي كالارقام والحروف والرموز وغيرها		