# Liste des acronymes

KB knowledge base.

NER Named Entity Recognition.

NLP Natural Language Processing.

TALN Traitement automatique du language natuelle.

IA Intelligence artificielle.

RE Relation Extraction.

# Introduction Générale :

### **Contexte**

Dans les agences de maintien de l'ordre, la collecte et l'analyse précises des déclarations des suspects, des témoins et des victimes sont cruciales pour mener à bien les enquêtes criminelles. Les entretiens et les interrogatoires jouent un rôle central dans ces enquêtes, fournissant des informations essentielles qui peuvent déterminer la direction de l'enquête et influencer les décisions judiciaires. Cependant, la gestion de ces informations présente de nombreux défis, notamment la détection des contradictions dans les déclarations et la formulation de questions pertinentes pour obtenir des informations supplémentaires.

Traditionnellement, les enquêteurs prennent des notes manuelles ou enregistrent des conversations, ce qui nécessite un effort considérable pour analyser et comparer les informations recueillies. Cette approche peut être sujette à des erreurs humaines et à des pertes d'informations importantes. De plus, les enquêteurs peuvent rencontrer des difficultés lorsqu'ils sont confrontés à des déclarations ambiguës ou incohérentes, et lorsqu'ils doivent formuler des questions de suivi pour clarifier ou approfondir les réponses obtenues.

Avec l'avènement des technologies d'intelligence artificielle (IA) et du traitement automatique du langage naturel (TALN), il est désormais possible de développer des systèmes automatisés capables d'analyser les conversations de manière plus efficace et précise. Ces technologies peuvent aider à détecter les contradictions dans les déclarations, en comparant les nouvelles informations avec les données existantes dans le système. Elles peuvent également fournir des recommandations de questions basées sur le contexte de l'entretien, aidant ainsi les enquêteurs à obtenir des informations plus complètes et cohérentes.

Des systèmes similaires ont été développés dans divers domaines du maintien de l'ordre, tels que la détection de la fraude, la reconnaissance faciale et l'analyse prédictive du crime[[1]](#footnote-1) [[2]](#footnote-2). Ces systèmes montrent le potentiel de l'IA à transformer les pratiques d'enquête en automatisant des tâches complexes et en fournissant des analyses avancées. Cependant, un système spécifiquement conçu pour enregistrer les conversations d'enquête, détecter les contradictions et recommander des questions en temps réel représente une avancée significative dans ce domaine.

**Contribution**

Ce projet vise à combler cette lacune en fournissant un outil intégré et automatisé pour les enquêteurs. En exploitant les capacités des technologies d'IA et de TALN, le système proposé améliorera la précision et l'efficacité des enquêtes, tout en réduisant la charge cognitive des enquêteurs. Cette initiative s'inscrit dans une tendance plus large d'adoption de technologies avancées par les agences de maintien de l'ordre pour améliorer leurs capacités d'investigation et garantir une justice plus efficace et équitable.

**Structure de la mémoire**

Cette thèse est structurée pour fournir une exploration complète du développement et de la mise en œuvre d'un système basé sur l'IA pour l'analyse des entretiens des forces de l'ordre. Les chapitres suivants décrivent la progression de cette entreprise de recherche, abordant chacun des aspects cruciaux de la conception du système, de ses fonctionnalités, de son évaluation et de ses implications futures. La structure détaillée est la suivante :

**Chapitre 1 : Présentation de l’organisme d’accueil et Généralités**

* Dans ce chapitre, nous présenterons brièvement l'organisme d'accueil, son rôle et sa mission dans le contexte plus large de son domaine d'activité. Nous aborderons également quelques généralités sur le domaine.

**Chapitre 2 : État de l’art**

* Ce chapitre offre une exploration approfondie de la littérature pertinente, comprenant une introduction au Traitement du Langage Naturel (NLP), ses différentes tâches, ainsi qu'une discussion sur les lacunes de la recherche actuelle.

**Chapitre 3 : Conception**

* Où nous décrirons en détail la conception de notre contribution et de notre application.

**Chapitre 4 : Implémentation et évaluations**

* Où nous aborderons les aspects techniques de la mise en œuvre de notre système, ainsi que les évaluations réalisées.

**Conclusion générale**

* La thèse se termine par un résumé des contributions, des discussions sur les limitations et des suggestions pour les orientations futures de la recherche, encapsulant la signification et le potentiel du système développé.

Table of Contents

[Introduction Générale : 1](#_Toc167600275)

[**Contexte** 1](#_Toc167600276)

[Chapitre 1 : Présentation de l’organisme d’accueil et Généralités 5](#_Toc167600277)

[1 Présentation de l’organisme d’accueil 5](#_Toc167600278)

[1.1.1 Introduction 5](#_Toc167600279)

[1.1.2 La Gendarmerie Nationale 5](#_Toc167600280)

[1.1.3 INCC-GN 5](#_Toc167600281)

[1.2 Généralités 6](#_Toc167600282)

[1.2.1 Introduction 6](#_Toc167600283)

[1.2.2 Crime : 6](#_Toc167600284)

[1.2.3 Enquête : 6](#_Toc167600285)

[1.2.4 Audition 7](#_Toc167600286)

[1.2.5 Contradiction 7](#_Toc167600287)

[1.2.6 Recommandation de questions 7](#_Toc167600288)

[Chapitre 2 : État de l’art 7](#_Toc167600289)

[2.1 Traitement automatique du langage naturel 7](#_Toc167600290)

[2.2 Représentation du texte 8](#_Toc167600291)

[2.2.1 Représentation textuelle discrète 8](#_Toc167600292)

[2.2.2 Représentation textuelle continue 9](#_Toc167600293)

[2.2.3 Réseaux de Neurones Récurrents 10](#_Toc167600294)

[2.3.4 LSTM: Long-Short-Term-Memory 12](#_Toc167600295)

[2.3.5 Transformers 13](#_Toc167600296)

[2.3.6 BERT 15](#_Toc167600297)

[2.3 Les différentes taches de NLP 17](#_Toc167600298)

[2.3.1 Reconnaissance d’entités nommées 17](#_Toc167600299)

[2.3.1.1 Définition 17](#_Toc167600300)

[2.3.1.2 Les approches du NER 17](#_Toc167600301)

[2.3.2 Classification de texte 18](#_Toc167600302)

[2.3.3 Extraction de relations 18](#_Toc167600303)

[2.3.4 Similarité Des Textes 19](#_Toc167600304)

[2.4 **Travaux Connexes** 19](#_Toc167600305)

[2.5 Recherches dans la Contradiction 20](#_Toc167600306)

[2.6 Systèmes de recommandation 20](#_Toc167600307)

[2.7 Recommandation des questions 21](#_Toc167600308)

[2.8 Filtrage basé sur le contenu (Content-based filtering) 21](#_Toc167600309)

[Chapitre 3 : Conception 21](#_Toc167600310)

[3.1 Introduction 22](#_Toc167600311)

[3.2 Diagramme de cas d’utilisation 22](#_Toc167600312)

[3.3 Diagramme de classes 22](#_Toc167600313)

[3.4 Module de détection de la contradiction 22](#_Toc167600314)

[3.5 Module de recommandation des questions 23](#_Toc167600315)

# Chapitre 1 : Présentation de l’organisme d’accueil et Généralités

## 1 Présentation de l’organisme d’accueil

### 1.1.1 Introduction

Durant cette partie, nous présenterons l’organisme d’accueil au sein de laquelle nous avons effectué notre projet, nous parlerons d’abord de la gendarmerie nationale ensuite de L’institut Nationale de Criminalistique et de Criminologie et la relation entre les deux et nous finirons par aborder les besoins de notre organisme d’accueil.

### 1.1.2 La Gendarmerie Nationale

La Gendarmerie nationale d'Algérie est une force de sécurité chargée de maintenir l'ordre public, d'assurer la sécurité des citoyens et de défendre l'intégrité du territoire national. Elle opère sous l'autorité du ministère de la Défense nationale et joue un rôle crucial dans la protection des populations, la lutte contre la criminalité, le contrôle des frontières et la gestion des situations d'urgence. La Gendarmerie nationale algérienne est reconnue pour son professionnalisme, son engagement et son dévouement envers la sécurité et le bien-être des citoyens.

### 1.1.3 INCC-GN

L'Institut National de Criminalistique et de Criminologie de la Gendarmerie Nationale (INCC-GN) est une institution algérienne spécialisée dans la recherche, l'analyse et la résolution des affaires criminelles. Relevant de la Gendarmerie nationale, cet institut joue un rôle essentiel dans la lutte contre la criminalité en fournissant des expertises scientifiques et techniques aux enquêteurs sur le terrain. L'INCC-GN est doté de laboratoires modernes équipés des technologies les plus avancées, permettant l'analyse des preuves et des indices recueillis lors des enquêtes. Son personnel hautement qualifié, composé d'experts en criminalistique et en criminologie, travailles-en étroite collaboration avec les autorités judiciaires pour résoudre les affaires criminelles, apporter des éléments de preuve solides devant les tribunaux et contribuer à la sécurité publique. Ces missions incluent :

* Réaliser des expertises et examens scientifiques à la requête des magistrats, enquêteurs et autorités habilitées.
* Assurer une assistance scientifique aux unités lors des investigations complexes.
* Concevoir et réaliser des banques de données, conformément à la loi.
* Participer aux études et analyses relatives à la prévention et à la réduction de toute forme de criminalité.
* Contribuer à la définition d'une meilleure politique de lutte contre la criminalité.
* Initier et mener des travaux de recherche ayant trait à la criminalité en recourant à des technologies de pointe.
* Œuvrer au développement de la recherche appliquée et des méthodes d'investigation ayant été jugées efficaces dans les domaines de la criminologie et de la criminalistique sur les plans national et international.[[3]](#footnote-3)

Nous contribuons avec ce projet aux derniers objectifs mentionnés ci-dessus.



Figure 1.1 : Logo de La Gendarmerie Nationale

## 1.2 Généralités

### 1.2.1 Introduction

Cette partie est un bref aperçu des concepts auxquels notre étude s’intéresse notamment l’audition, contradiction, et la recommandation de questions. Les sections suivantes rassemblent quelques définitions et éléments préliminaires qui sont nécessaires pour la bonne compréhension du contenu du mémoire.

### 1.2.2 Crime :

Un crime, dans son essence la plus fondamentale, est un acte qui viole les lois établies par une société donnée. C'est un comportement considéré comme répréhensible et interdit par la loi, et qui entraîne des conséquences juridiques pour son auteur. Les crimes peuvent varier en gravité, allant des infractions mineures telles que le vol à l'étalage aux offenses les plus graves comme le meurtre ou le viol. Un crime qui a été découvert déclenche une enquête.

### 1.2.3 Enquête :

Une enquête est un processus systématique et méthodique visant à recueillir des informations, à examiner des preuves et à analyser des faits dans le but de découvrir la vérité sur un événement spécifique ou une série d'incidents. C'est une démarche essentielle dans le domaine de la justice et de la résolution de problèmes, utilisée par les autorités légales, les organismes d'application de la loi, les entreprises et d'autres entités pour élucider des questions complexes et prendre des décisions éclairées.

Le processus d'enquête implique généralement plusieurs étapes, notamment la collecte initiale d'informations, la planification de la méthode d'investigation, la collecte de preuves, l'analyse des données recueillies, la formulation de conclusions et, le cas échéant, la présentation des résultats devant un tribunal ou une autre instance compétente.

### 1.2.4 Audition

L'audition joue un rôle crucial dans le domaine de la sécurité publique, notamment dans le cadre des enquêtes judiciaires et des interventions policières. Elle permet de recueillir des informations essentielles auprès des témoins, des suspects et des victimes, contribuant ainsi à la résolution des affaires criminelles et à la protection des citoyens.

Les professionnels de la sécurité publique, tels que les officiers de police judiciaire, doivent posséder des compétences avancées en techniques d'interrogatoire pour obtenir des déclarations précises et fiables. L'audition efficace repose sur la capacité à poser des questions pertinentes, à détecter les contradictions et à évaluer la crédibilité des informations fournies.

### 1.2.5 Contradiction

Une contradiction se produit lorsqu'il existe une incohérence ou un conflit entre deux assertions, déclarations ou faits. C'est un phénomène qui peut survenir dans divers contextes, que ce soit dans des déclarations verbales, des documents écrits, des témoignages ou des arguments logiques. La détection et la compréhension des contradictions sont essentielles dans de nombreux domaines, notamment le droit, la recherche, la communication et la résolution de problèmes.

Dans un contexte juridique, la contradiction peut être utilisée pour remettre en question la crédibilité d'un témoignage ou d'une preuve présentée devant un tribunal. Les enquêteurs peuvent chercher à identifier des contradictions dans les déclarations des témoins ou les éléments de preuve présentés par la partie adverse afin de discréditer leur argumentation et de renforcer leur propre cas.

### 1.2.6 Recommandation de questions

La recommandation de questions est un processus visant à proposer des interrogations pertinentes et efficaces dans le cadre d'une enquête, d'un entretien ou d'une interaction. Cette pratique trouve son utilité dans divers domaines, notamment dans le cadre juridique, médical, professionnel et académique.

Dans le domaine juridique, la recommandation de questions est souvent utilisée lors des interrogatoires de témoins ou de suspects. Les enquêteurs peuvent bénéficier de recommandations basées sur des analyses de cas similaires, des modèles de comportement ou des profils psychologiques pour formuler des questions pertinentes et stratégiques afin de faire progresser l'enquête et obtenir des informations cruciales.

# Chapitre 2 : État de l’art

Dans ce chapitre, nous allons présenter les concepts de base liés à notre travail,

## 2.1 Traitement automatique du langage naturel

Le traitement automatique du langage naturel (TALN), ou Natural Language Processing (NLP) en anglais, est un domaine essentiel de l'intelligence artificielle qui vise à permettre aux machines de comprendre et d'interpréter le langage humain. Cette discipline repose sur la combinaison de la linguistique informatique, des techniques statistiques, et des algorithmes d'apprentissage automatique et d'apprentissage profond[[4]](#footnote-4).

Les technologies NLP permettent aux ordinateurs de traiter le texte et les données vocales de manière à extraire des informations significatives, comme le contexte, l'intention et les émotions[[5]](#footnote-5). Grâce à ces capacités, le NLP trouve des applications variées dans de nombreux domaines.

Par exemple, les systèmes de traduction automatique, les assistants virtuels comme Siri et Alexa, et les chatbots utilisent des techniques de NLP pour interagir avec les utilisateurs de manière fluide et naturelle. De plus, les entreprises emploient le NLP pour analyser les avis des clients, détecter le spam dans les emails, et identifier les entités nommées dans les documents.

Le traitement du langage naturel joue un rôle crucial dans l'avancement de diverses technologies contemporaines. Voici quelques exemples illustratifs :

• Modèles de langage comme GPT, chatbots

• Reconnaissances vocales

• Traduction automatique, analyse des sentiments

## 2.2 Représentation du texte

Pour que les machines puissent comprendre et analyser les modèles linguistiques, il est essentiel de convertir les mots en nombres. Ce processus, appelé représentation textuelle, est fondamental pour la plupart des tâches de traitement du langage naturel (NLP). La manière dont le texte est représenté influence grandement les performances des modèles d'apprentissage automatique. On distingue deux grandes catégories de représentations textuelles :

* Représentation textuelle discrète
* Représentation textuelle continue

### 2.2.1 Représentation textuelle discrète

Dans ce type de représentation, les mots sont représentés par des indices correspondant à leur position dans un dictionnaire dérivé d'un corpus plus large. Les méthodes suivantes sont couramment utilisées :

* One-Hot encoding
* Bag of Words
* CountVectorizer
* TF-IDF

#### 2.2.1.1 One-Hot encoding

Cette méthode représente chaque mot par un vecteur binaire de la taille du vocabulaire, avec un seul élément à 1 et tous les autres à 0.

Exemple : Pour le vocabulaire ["chat", "chien", "oiseau"], "chien" est représenté par [0, 1, 0].

Avantage : Simple à comprendre et implémenter.

Limite : Produit des vecteurs de grande taille pour les grands vocabulaires et ne capture pas les relations sémantiques entre les mots.

#### 2.2.1.2 Bag of words

La représentation par sac de mots, comme son nom l'indique intuitivement, place les mots dans un « sac » et calcule la fréquence d'apparition de chaque mot. Elle ne prend pas en compte l'ordre des mots ou les informations lexicales pour la représentation du texte.

L'intuition derrière la représentation par sac de mots est que des documents ayant des mots similaires sont similaires, indépendamment de la position des mots.

Avantage : Simple à implémenter et efficace pour les petites tâches.

Limite : Ne capture pas l'ordre des mots ni les relations sémantiques.

#### 2.2.1.3 CountVectorizer

Le CountVectorizer est une extension de BoW, où chaque document est représenté par un vecteur contenant le nombre de fois que chaque mot apparaît.

Exemple : Pour les phrases "le chat dort" et "le chien dort", "le chat dort" pourrait être représenté par [1, 1, 1, 0] en utilisant le même vocabulaire.

Avantage : Simple et intuitif, permet de quantifier l'importance des mots.

Limite : Ne prend pas en compte la fréquence relative des mots dans l'ensemble des documents.

#### 2.2.1.1 TF-IDF

Le TF-IDF (Term Frequency-Inverse Document Frequency) pondère chaque mot en fonction de sa fréquence dans le document et de sa rareté dans l'ensemble des documents. Il combine deux mesures pour évaluer l'importance d'un mot dans un document :

* **TF (Term Frequency)** : la fréquence d'apparition d'un mot **m** dans un document **d**.
* **IDF (Inverse Document Frequency)** : la mesure de la rareté d'un mot dans l'ensemble des documents, calculée comme suit :

Où **N** est le nombre total de documents et **df(m)** le nombre de documents contenant le mot **m**.

Avantage : Réduit l'impact des mots courants et met en valeur les mots informatifs.

Limite : Les vecteurs restent de grande dimension et la méthode ne capture pas les relations contextuelles entre les mots.

### 2.2.2 Représentation textuelle continue

La représentation continue, ou distribuée, d'un texte se produit lorsque la représentation d'un mot dépend d'autres mots et n'est pas mutuellement exclusive. Les configurations des mots reflètent diverses métriques et concepts présents dans les données. Ainsi, les informations relatives à un mot sont réparties le long du vecteur qui le représente. Cela contraste avec la représentation discrète, où chaque mot est considéré comme unique et indépendant des autres.

Les méthodes courantes de représentation continue incluent :

* Matrices de cooccurrence
* Word2Vec
* GloVe

#### 2.2.2.1 Matrice de cooccurrence :

#### Cette représentation analyse la proximité des entités entre elles au sein d'un texte. Une entité peut être un seul mot, un bi-gramme (séquence de deux mots), ou même une phrase, bien que l'utilisation d'un seul mot soit la méthode la plus courante pour calculer la matrice. En examinant les co-occurrences, cette matrice permet de dévoiler les associations et les relations entre différents mots dans un corpus. Cela nous aide à comprendre comment les mots sont utilisés ensemble et à identifier les tendances et les motifs linguistiques présents dans le texte.

#### 2.2.2.2 Word2Vec :

Word2Vec est un algorithme de word embedding, qui représente les mots ou phrases d’un texte sous forme de vecteurs de nombres réels dans un modèle vectoriel. Développé par une équipe de recherche de Google sous la direction de Tomas Mikolov[[6]](#footnote-6), cet algorithme est devenu un outil essentiel dans le traitement du langage naturel.

Word2Vec propose deux architectures neuronales principales : CBOW (Continuous Bag of Words) et Skip-Gram.

* **CBOW (Continuous Bag of Words)** : Cette méthode tente de prédire un mot en se basant sur son contexte, c’est-à-dire les termes qui l’entourent dans une phrase. Elle est particulièrement efficace avec des ensembles de données plus petits et offre un temps d'entraînement rapide comparé à Skip-Gram.
* **Skip-Gram** : Cette méthode, à l'inverse de CBOW, prédit le contexte à partir du mot cible. Elle tend à mieux fonctionner avec des ensembles de données plus larges, bien qu’elle nécessite un temps d'entraînement plus long.

#### 2.2.2.3 GloVe :

GloVe est un autre algorithme d’embedding[[7]](#footnote-7), il essaie d’englober le meilleur de la technique basée sur le comptage (matrice de cooccurrence) et de la technique basée sur la prédiction (Word2Vec), d’où viens également l’appellation de technique hybride.

Dans cet algorithme, nous essayons d’appliquer la relation ci-dessous :

viT vj = logP(i|j)

Où :

• Vi est la transposée de la iéme representation du mot.

• Vj est la représentation du jéme mot.

• P(i|j) est la probabilité que le mot j apparaisse dans le contexte du mot i.

Donc, essentiellement, nous construisons des vecteurs de mots Vi et Vj pour être fidèles à P(i|j) qui est une statistique calculée globalement à partir de la matrice de cooccurrence.

#### 2.2.2.4 Transformers :

Voir section 2.3.5

### 2.2.3 Réseaux de Neurones Récurrents

Les réseaux de neurones récurrents (RNN) sont un type de réseau de neurones qui est utilisé pour le traitement des données séquentielles tel que du texte, de l’audio, etc... Ce qui les distingue des autres réseaux de neurones est leurs "mémoire", en effet, là ou un réseau de neurones classique traite chaque donnée indépendamment, un réseau de neurones récurrent prend en considération les états précédents, à chaque passage de donnée vers le réseau il y a une concaténation des données en entrée ainsi que de l’état précédent, ainsi cela permet au réseau de d’apprendre et de comprendre le contexte d’une phrase ou d’un audio pour ainsi pouvoir par exemple prédire le prochain mot. Concrètement, un réseau de neurones récurrent se présente comme suit :

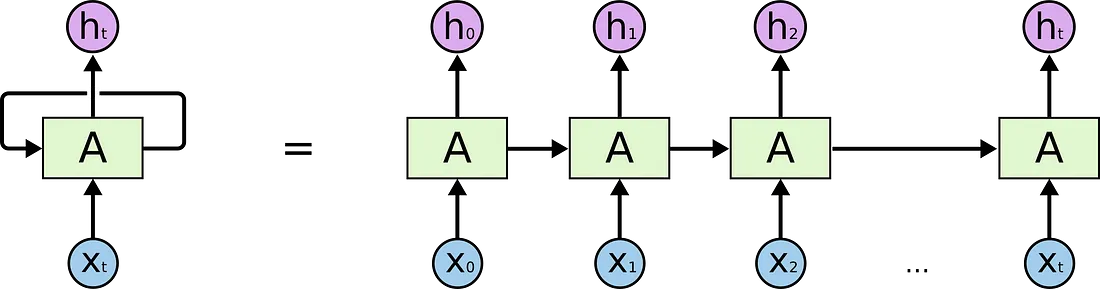


Figure 2 - source : https://medium.com/@jianqiangma/all-about-recurrent-neural-networks-9e5ae2936f6e

L’élément clé de ce type de réseau est la structure de ces cellules (perceptron) ainsi que la manière de concaténer le vecteur représentant l’état précèdent et les données en entrée.

**La mémoire** Nous avons évoqué un peu plus haut le concept de mémoire comme étant le fait qu’un RNN prend en considération les états précédents, plus concrètement comme on peut le voir dans la figure 2.3 un vecteur hi i représentant le numéro de l’itération, est responsable de cette opération.

Le problème avec une telle architecture, malgré son efficacité a traité les données séquentielles, sont nombreux, parmi eux nous citerons

• Problème de mémoire à long terme

Le problème de mémoire à long terme peut être illustré par l’exemple suivant : Supposons que nous voulions prédire le prochain mot dans une phrase, nous donnerons en entrée à notre réseau chaque mot représenté sous forme vectorielle (voir section 2.3) admettons la phrase suivante en entrée :

" Unix est un système d’exploitation, il fut créé par Ken Thomson et Dennis Richie, ces deux personnages sont des légendes dans le domaine de l’informatique très peu reconnu par la nouvelle génération sans eux le monde ne serait peut-être pas celui qu’il est aujourd’hui, Linux un autre système d’exploitation apparu en 1991 est basé sur \*\*\*\* "

Afin de prédire Unix le modèle devrait prendre en considération (entre autre) le fait qu’Unix est un système d’exploitation, or cette information est assez lointaine au niveau de la phrase, un réseau de neurones récurrent classique aura donc de grande difficulté pour les données qui nécessite une mémoire à long terme.

• Problème de parallélisation

Un RNN traitant les données de manière séquentielle, il est impossible d’utiliser toutes les optimisations de matérielle et logicielle afin de paralléliser et améliorer la vitesse d’entraînement d’un tel réseau, à cause de son caractère séquentielle.

Plusieurs solutions existent afin de régler les différents problèmes que rencontre les RNN classique, parmi elle l’utilisation d’unité récurrente (cellule) plus complexe dite "gated cells" afin de mieux contrôler le flux de donnée, l’idée étant de filtrer les informations et ne garder que la plus pertinente. Parmi ces réseaux réseau, il y a notamment les LSTM, GRU..

### 2.3.4 LSTM: Long-Short-Term-Memory

Comme mentionné plus haut, les LSTM ont été créé afin de palier au problème de mémoire a long terme (d’où leur nom d’ailleurs) grâce à l’utilisation de cellule mémoire plus complexe nommée "gated cells". Ces cellules font office de contrôleurs de flux de données, elles sont composées de 3 entrées/sortie et de 3 portes (gates) à l’intérieur cellule comme le montre la figure 2.4.

Deux des entrées/sortie "Cell state" et "Hidden state" contiennent des valeurs issues de l’état précédent, la 3e sert d’entrée à l’état actuelle (pour un texte c’est un mot d’une phrase ou d’un paragraphe) la première est censée encoder l’agrégation des données de tous les états précédents, la seconde a pour but d’encoder la caractérisa- tion1 de l’entrée précédente.

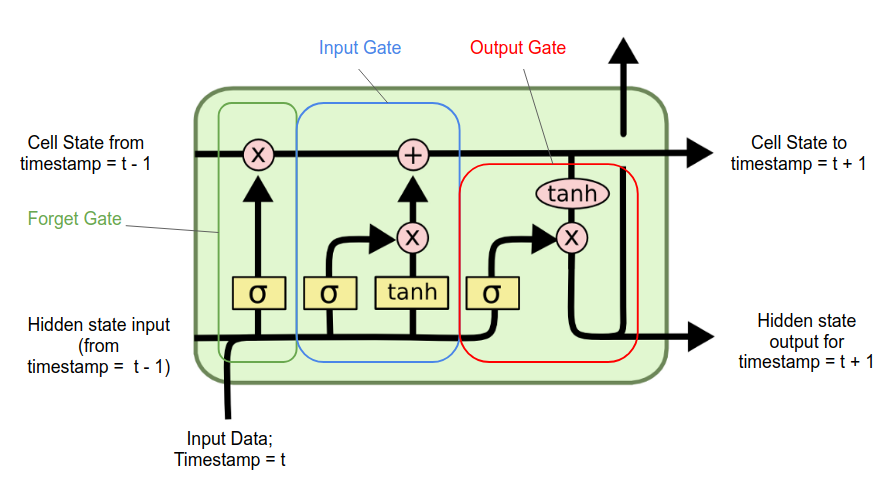


Figure 3 - Structure d’une cellule de LSTM

Pour ce qui est des 3 portes, la première "forgot" (oublie) permet au réseau de diminuer (et donc aussi accroître) l’importance de certains mots par rapport à d’autres notamment grâce à la fonction sigmoïde suivi de la multiplication "pointwise" (point par point) qui permet d’amplifier ou/et de rabaissé les valeurs. Tel que le montre l’équation (2.1) et la figure (4)

• ft sortie de la forgot gate.

• Wf poids de la forgot gate.

• ht−1 résultats (sortie) du précédent LSTM (a l’étape t-1).

• xt entrée a l’état présent.

• bf biais de la porte oublie.

La porte entrée "input" elle, se charge d’extraire les caractérisations significatives pour le LSTM (significatif par rapport à notre besoin). La dernière porte, celle de la sortie "output" sert a combiner les résultats des précédentes portes afin de former la sortie "hidden" qui peut être utilisé soit comme prochaine entrée ht ou pour effectuer une prédiction yt. (voir figure 4)

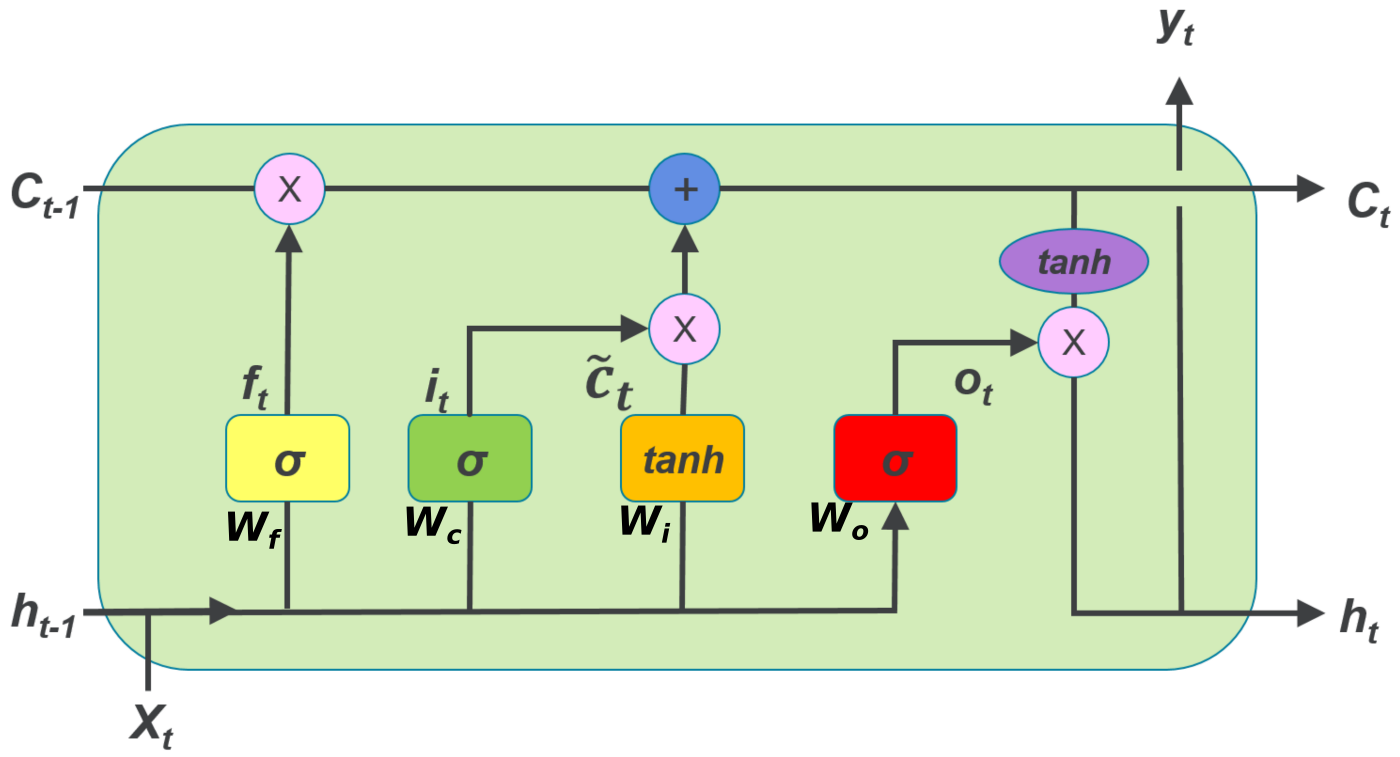


Figure 4 - Exemple d’une structure standard d’un réseau de neurones récurrent de type LSTM avec annotations

### 2.3.5 Transformers

Le transformer est un modèle d’apprentissage profond introduit en 2017 dans l’article "Attention is all you need"[[8]](#footnote-8) qui utilise le mécanisme d’attention avec un système d’encodeur et de décodeur. Cette architecture a été proposée par les cherchers de Google Brain et Google Resarch, dans le but de surmonter le défaut de la structure LSTM, qui est l’entraînement lent du modèle.

En effet, grâce au mécanisme d’attention, on peut identifier le contexte de notre input, qui nous permettra d’utiliser la parallélisation, ce qui réduira considérablement le temps d’entraînement. Ci-dessous, nous montrons l’architecture du transformer :

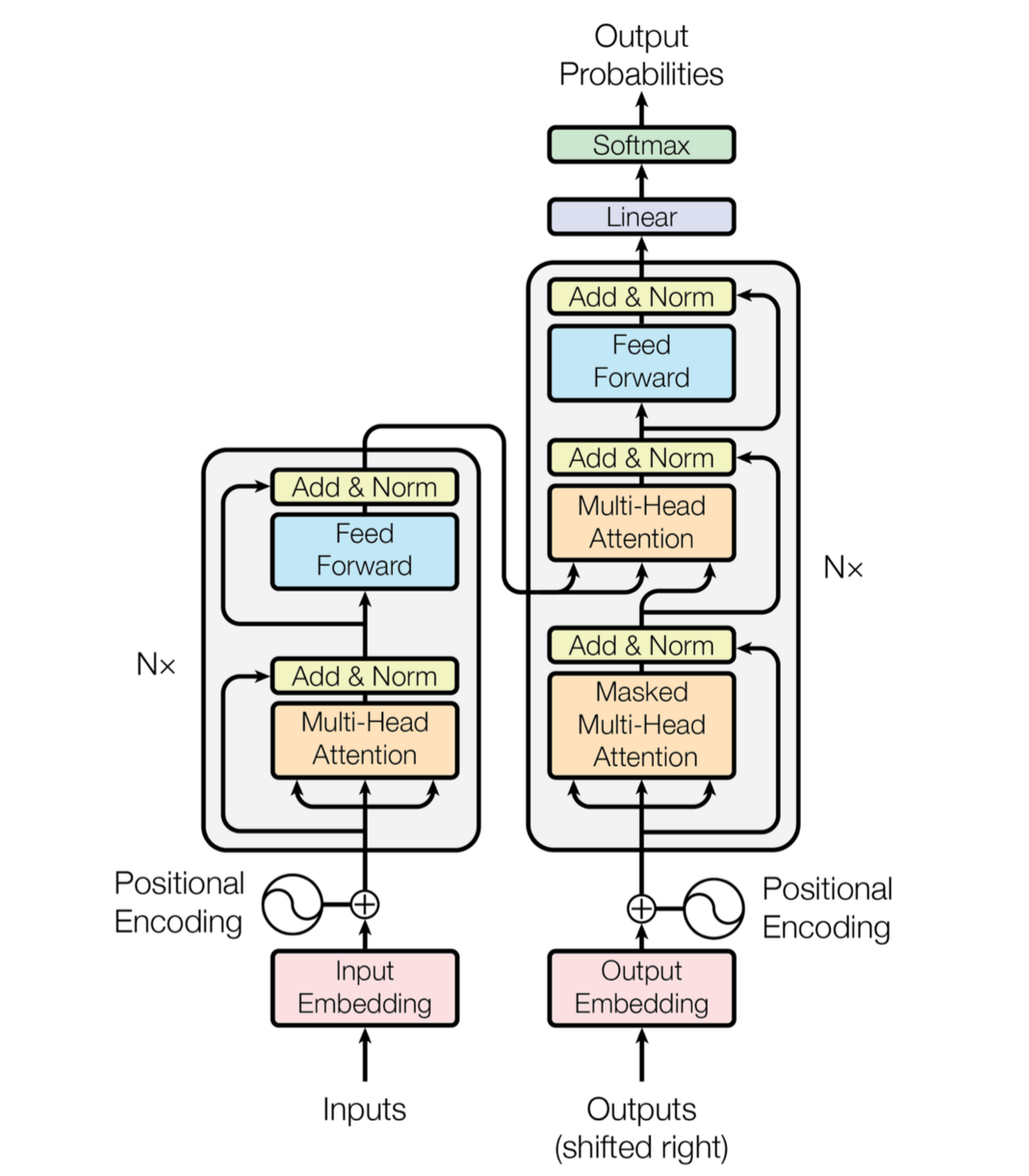


Figure 5 - L’architecture du transformer

Nous allons maintenant expliquer les mécanismes clé qui forme le transformer :

#### 2.3.5.1 Mécanisme d’attention

Le mécanisme de self-attention détermine l’interdépendance des différents mots de la même séquence afin de lui associer une représentation (encoding) pertinente, ce qui donne à notre input plus d’informations contextuelles.

Le vecteur de self-attention est calculé depuis les vecteurs suivant :

• Vecteur requête (Query Vector) qu’on appelle q.

• Vecteur clé (Key Vector) qu’on appelle k.

• Vecteur valeur (Value Vector) qu’on appelle v.

On indice chacun de ces vecteurs par la position du mot dans la séquence (Ex: q1,k1 et v1 pour le premier mot de la séquence). Ces vecteurs sont obtenus en multipliant l’embedding de la séquence d’entrée par les 3 matrices qui sont entraînées pendant le processus d’entraînement du Transformer.

#### 2.3.5.2 Encodeur

Depuis l’article Attention is all you need: "L’encodeur est composé d’un empilement de N = 6 couches identiques. Chaque couche a deux sous-couches. Le premier est un mécanisme de multi-head self-attention, et le second est un réseau de feed-forward."

L’entrée d’un encodeur est la sortie du précédent. L’entrée du premier encodeur est un vecteur d’embedding, la sortie du dernier encodeur est ensuite utilisé dans le décodeur.

#### 2.3.5.3 Décodeur

Comme l’encodeur, le décodeur est un empilement de 6 couches identiques, qui contient une sous-couche de self-attention et un réseau feed-forward, mais il contient en plus une couche « Encoder-Decoder Attention » qui a pour but de permettre au décodeur de réaliser le mécanisme d’attention entre la séquence d’entrée (encodée) et la séquence de sortie (en train d’être décodée).

Le dernier décodeur est connecté à un bloc Réseau de neurones linéaire + Soft-max. Le rôle de ce bloc est de permettre d’identifier à quels mots du vocabulaire correspondent les sorties du dernier encodeur.

### 2.3.6 BERT

#### 2.3.6.1 Introduction

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de transformer développé par une équipe de Google en 2018[[9]](#footnote-9), BERT est conçu pour pré-entraîner des représentations bidirectionnelles profondes à partir de texte non labellisé en conditionnant conjointement les contextes gauche et droite dans le but d’avoir un sens plus profond du contexte linguistique.

Avec un réglage-fin (fine-tuning) du modèle, qui consiste à ajouter une seule couche de sortie supplémentaire, BERT peut créer des résultats à la pointe de la technologie. Cela est possible grâce à la technique MLM (named Mask LM) qui permet un entraînement bidirectionnel dans des modèles dans lesquels il était auparavant impossible.

Ci-dessous les procédures globales de pre-training et de fine-tuning pour BERT :

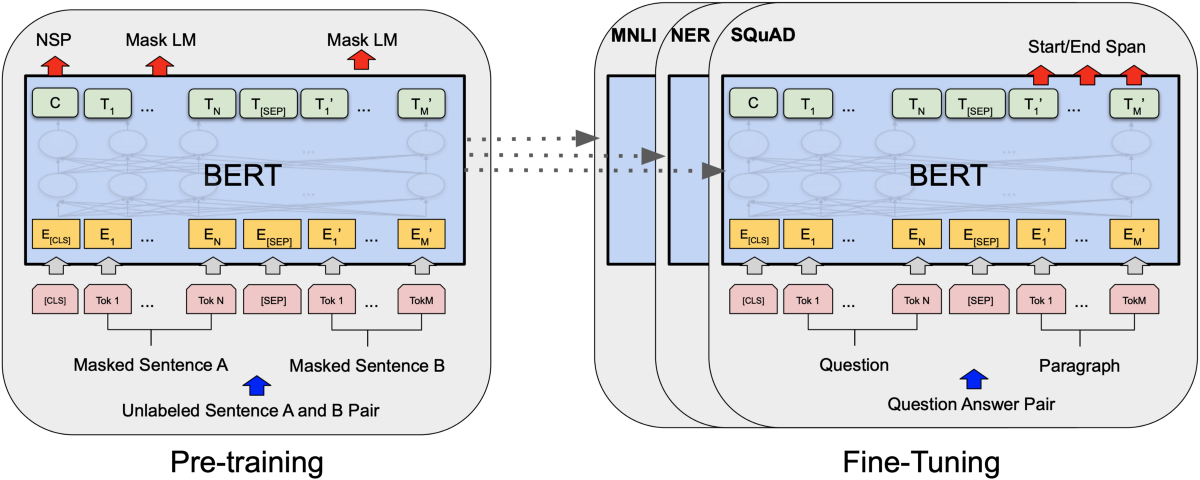


Figure 6 - L’architecture de BERT pour le pre-training et le Fine-Tuning

#### 2.3.6.2 Architecture de BERT

BERT est un encodeur de transformateur bidirectionnel multicouche, deux modèles de BERT ont été conçu : le modèle de base, et le modèle large. La différence réside dans le nombre de couches, taille cachée et le nombre de têtes de self-attention.

La représentation de données d’entrées peut être à la fois une seule phrase et une paire de phrases dans une seule séquence de token, ce qui nous aidera dans l’implémentation des taches nlp sur des données conversationnelles.

#### 2.3.6.3 Pre-training de BERT

Nous allons expliquer deux concepts clés qui ont rendu BERT le modèle standard pour le transfer-learning :

• Masked LM :

Pour chaque séquence, on remplace un pourcentage de mots par le token [MASK], puis le modèle essaie de les prédire en fonction du contexte fourni par les autres mots non masqués de la séquence.

• Next Sentence Prediction (NSP) :

Dans le processus d’entrainement de BERT, le modèle reçoit des paires de phrases en entrée et apprend à prédire si la deuxième phrase de la paire est la phrase suivante dans le document d’origine. Pendant l’apprentissage, 50 % des entrées sont une paire dans laquelle la deuxième phrase est la phrase suivante dans le document d’origine, tandis que dans les 50 % restants, une phrase aléatoire du corpus est choisie comme deuxième phrase.

Ce concept est utilisé pour le but de comprendre la relation entre deux phrases, qui est important pour des taches NLP comme Réponse aux questions et inférence en langage naturel.

#### 2.3.6.4 Fine-tuning de BERT

Le fine-tuning permet à BERT de modéliser plusieurs taches NLP -tout en étant relativement peu couteux-, en ajoutant seulement une couche au modèle de base :

• Les tâches de classification telles que l’analyse des sentiments est effectuée en ajoutant une couche de classification au-dessus de la sortie Transformer pour le token [CLS].

• Pour la reconnaissance d’entité nommée, le modèle reçoit une séquence de texte et doit marquer les différents types d’entités qui apparaissent dans le texte. Un modèle NER peut être formé en alimentant le vecteur de sortie de chaque token dans une couche de classification qui prédit l’étiquette.

## 2.3 Les différentes taches de NLP

Les subtilités du langage humain rendent extrêmement complexe la tâche de développer des logiciels capables de comprendre avec précision le sens voulu des textes ou d’autres données. Des phénomènes tels que les homonymes, les homophones, le sarcasme, les métaphores et les variations de la structure des phrases ne sont que quelques exemples des défis que présente le langage humain. Ces subtilités, qui peuvent prendre des années à être maîtrisées par les humains, doivent être apprises par les programmeurs dès le début pour que les applications basées sur le langage naturel puissent reconnaître et comprendre correctement ces nuances, et ainsi être véritablement utiles.

Dans le domaine du traitement automatique du langage naturel (TALN), plusieurs tâches sont définies pour aider les ordinateurs à interpréter les données textuelles. Voici quelques exemples de ces tâches qui nous utilisons dans notre projet :

### 2.3.1 Reconnaissance d’entités nommées

### 2.3.1.1 Définition

La reconnaissance d'entités nommées (Named Entity Recognition ou NER en anglais) est une sous-tâche de l'extraction d'informations visant à identifier et classer les mots clés, appelés entités, présents dans un document. Cette technologie permet de regrouper ces entités en catégories prédéfinies. Par exemple, dans un texte, la NER peut détecter et distinguer des mentions de personnes et de lieux, qui appartiennent à des catégories distinctes.

### 2.3.1.2 Les approches du NER

Comme évoqué précédemment, la reconnaissance d'entités nommées a pour but de repérer et de classer des mots dans un document textuel. Pour réaliser cette tâche, trois méthodes principales sont utilisées : l'approche statistique, l'approche basée sur des règles et l'approche hybride, qui combine les deux premières.[[10]](#footnote-10)

#### 2.3.1.2.1 Approche basé sur les règles

Cette méthode repose sur la définition d'un ensemble de règles grammaticales, syntaxiques, etc., établies par des linguistes, ainsi que sur l'utilisation de dictionnaires. Elle implique l'analyse du texte fourni en entrée pour identifier les entités et leurs catégories en appliquant les règles prédéfinies. Bien que cette approche offre une grande précision, elle exige un investissement considérable en termes de travail humain.

#### 2.3.1.2.2 Approche basée statistique

Cette méthode, contrairement à la précédente, s'appuie sur des règles statistiques et logiques pour atteindre le même objectif, en utilisant diverses techniques d'apprentissage automatique.

* **Apprentissage supervisé** : Cette méthode nécessite un corpus annoté pour entraîner le modèle. Le principal inconvénient est qu'elle ne se généralise pas bien, en raison du manque de grands ensembles de données pour ce type de tâche, ce qui nous pousse à envisager d'autres approches.
* **Apprentissage semi-supervisé** : Cette approche commence avec un corpus où seules quelques données sont étiquetées. Le modèle est d'abord entraîné sur ces données, puis le processus est itéré pour détecter d'autres entités similaires aux premières. Ce processus est répété en utilisant les résultats précédents pour affiner le modèle.
* **Apprentissage non supervisé** : Cette technique utilise le clustering, qui consiste à regrouper des éléments apparaissant dans des contextes similaires. Pour un texte donné, le modèle tente de trouver le groupe le plus similaire.

Malgré l'efficacité des méthodes d'apprentissage automatique et profond, le manque de grands ensembles de données les empêche d'atteindre la précision des approches basées sur des règles dans des domaines spécifiques. Cependant, ces méthodes offrent de meilleures performances en termes de généralisation.

#### 2.3.1.2.3 Approche Hybride

Dans cette approche, l'objectif est de fusionner les techniques basées sur des règles avec celles basées sur des statistiques pour tirer le meilleur parti des deux méthodes. L'idée est de parvenir à un compromis pour obtenir des résultats optimaux.

### 2.3.2 Classification de texte

La classification de texte est l'une des applications de TALN les plus largement utilisées en raison de ses nombreuses utilisations dans le monde réel telles que l'analyse des sentiments[[11]](#footnote-11) [[12]](#footnote-12) les critiques de produits[[13]](#footnote-13), la classification des actualités, ou la classification de l'intention de l'utilisateur dans les requêtes de recherche.

La classification de texte consiste à attribuer des catégories ou des étiquettes à des documents textuels en fonction de leur contenu. Cette technique permet de rendre les informations non structurées plus accessibles et exploitables. En utilisant des algorithmes sophistiqués, la classification de texte aide à automatiser et à accélérer le processus de tri et d'organisation de grandes quantités de données textuelles, ce qui est essentiel dans de nombreux domaines, de la recherche académique à l'industrie.

### 2.3.3 Extraction d'informations

L'extraction d'informations est la tâche de traitement du langage naturel (NLP) qui extrait des informations sémantiques structurées à partir de texte. Ces informations incluent des relations binaires - par exemple, des interactions biochimiques entre deux protéines[[14]](#footnote-14) ou des événements n-aires - c'est-à-dire des événements avec plus de deux arguments tels que des attaques terroristes, où chaque attaque est associée à plusieurs arguments, y compris l'emplacement de l'attaque, l'identité de l'attaquant, le nombre de victimes, le montant des dommages matériels, et ainsi de suite.[[15]](#footnote-15) L'extraction d'informations permet de nombreuses applications réelles importantes telles que la découverte de traitements potentiels pour les maladies ou la surveillance des attaques terroristes à partir de documents de presse.

### 2.3.4 Extraction de relations

### 2.3.4.1 Introduction

L'extraction de relations est une sous tache de l’extraction d’informations (IE) qui vise à identifier et à extraire les liens sémantiques entre différents éléments dans un texte. Ces éléments peuvent être des entités telles que des personnes, des lieux ou des événements, et les relations entre eux peuvent être diverses, allant des simples associations binaires aux structures plus complexes impliquant plusieurs entités. L'objectif principal de l'extraction de relations est de transformer le texte non structuré en données exploitables, facilitant ainsi la compréhension automatique des informations contenues dans les documents textuels. Cette tâche est cruciale dans de nombreux domaines, notamment la recherche d'informations, l'analyse des médias sociaux, la veille stratégique et la médecine.

### 2.3.4.2 Les approches du RE

Plusieurs approches existent pour l’extraction de relations[[16]](#footnote-16) :

#### 2.3.4.2.1 Approche basé sur les règles

Ces méthodes sont également appelées méthodes basées sur des patterns (motifs) construits manuellement. Ces types de méthodes définissent un ensemble de patterns d'extraction pour un ensemble prédéfini de relations. Ensuite, ces patterns d'extraction sont comparés au texte. Si un pattern correspond, une relation correspondant à ce pattern est trouvée dans le texte.

Par exemple un pattern pour les hyponymes comme ‘such X as Y’ avec le texte ‘such actors as angelina’ donne hyponyme(actor,angelina).

Les méthodes basées sur des règles nécessitent une expertise du domaine et des connaissances linguistiques pour définir des modèles d'extraction. Ces méthodes sont spécifiques à un domaine, reposant sur une structure de document fixe et des relations cibles prédéfinies. Lorsqu'on passe d'un domaine à un autre, il devient nécessaire de redéfinir les relations cibles et les modèles d'extraction. Par conséquent, les méthodes basées sur des règles demandent un effort manuel considérable et ne conviennent pas pour des corpus hétérogènes.

#### 2.3.4.2.2 Approche basé sur l’apprentissage non supervisé

Les méthodes non supervisées ne nécessitent pas de données annotées. La plupart des méthodes non supervisées d'extraction de relations utilisent une approche basée sur le clustering. L'une des premières approches non supervisées d'extraction de relations basée sur le clustering ont utilisé un étiqueteur d'entités nommées pour extraire les entités afin de se concentrer uniquement sur les relations avec les entités nommées mentionnées. Les étapes principales de la méthode d'apprentissage non supervisée sont :

1. Identification des entités nommées dans le corpus de texte
2. Identification des entités nommées co-occurrentes et de leur contexte
3. Regroupement des paires d'entités en fonction de la similarité de leur contexte
4. Attribution d'un nom de relation sémantique à chaque groupe.

#### 2.3.4.2.3 Approche basé sur l’apprentissage supervisé

Les méthodes supervisées nécessitent une grande quantité de données d'entraînement, annotées avec un ensemble d'entités et de relations. Elles utilisent ces données d'entraînement pour former un classificateur, qui extraira ensuite les relations des données de test. Il existe deux types de méthodes supervisées : les méthodes basées sur les caractéristiques et les méthodes basées sur les noyaux.

* Les méthodes basées sur les caractéristiques (features) :

Un ensemble de caractéristiques est généré pour chaque relation dans les données d'entraînement, puis un classificateur est entraîné à extraire une nouvelle instance de relation.

* Les méthodes basées sur les noyaux (kernal) :

Dans les approches basées sur les noyaux, des fonctions de noyau sont utilisées pour déterminer la similarité entre deux représentations d'instances de relation.

#### 2.3.4.2.3 Approche basé sur l’apprentissage semi-supervisé

La création de données pour les méthodes supervisées d'extraction de relations implique des coûts, des efforts et du temps. Cependant, les méthodes supervisées peuvent automatiser le processus de génération de données étiquetées grâce à des algorithmes de bootstrap. Cette approche offre deux avantages clés :

* Elle réduit l'effort nécessaire pour créer des données étiquetées
* Elle tire parti des données non étiquetées disponibles gratuitement.

L'algorithme de bootstrap repose sur une grande quantité de données non étiquetées et un petit ensemble d'instances de départ représentant le type de relation souhaité. Par exemple, pour extraire la relation "CapitaleDe", des exemples de départ comme (New Delhi, Inde), (Canberra, Australie) et (Londres, Angleterre) peuvent être utilisés pour développer un modèle d'extraction. Avec ces exemples de départ en entrée, l'algorithme de bootstrap est conçu pour identifier des relations similaires impliquant des paires d'entités telles que (Paris, France). La Figure 7 illustre le modèle pour extraire des motifs et des tuples de départ en utilisant une approche d'apprentissage semi-supervisé.

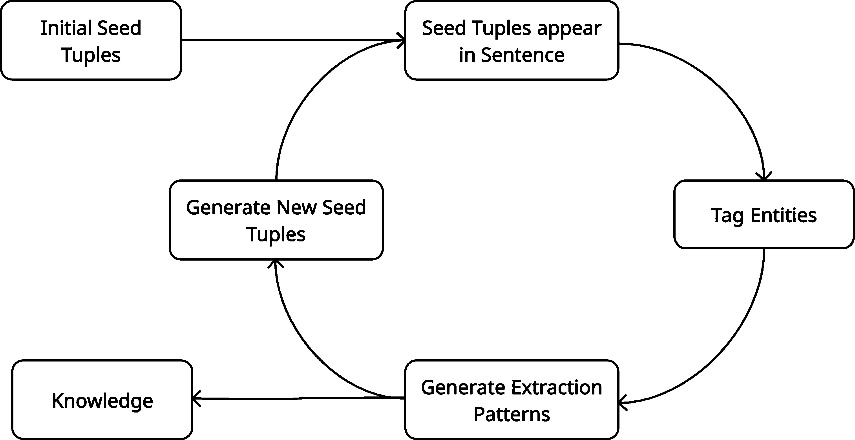


Figure 7 - Etapes pour l'apprentissage semi-supervisé de RE

#### 2.3.4.2.4 Approche de supervision distante

Ces méthodes sont également appelées méthodes supervisées faiblement ou basées sur la connaissance. Les chercheurs ont proposé une méthode dans laquelle les données d'entraînement sont automatiquement générées en alignant le texte avec une base de connaissances (KB), ce qui élimine le besoin d'étiquetage manuel. La supervision distante repose sur l'idée que si deux entités partagent une relation dans une KB, toutes les phrases mentionnant ces entités pourraient exprimer cette relation.

La supervision distante utilise une base de connaissances comme Freebase pour extraire les relations entre entités. Lorsqu'une paire d'entités apparaît à la fois dans une phrase et dans la KB, la phrase est liée de manière heuristique à la relation correspondante de la KB. Par exemple, dans la phrase "Bill Gates est le fondateur de Microsoft", si "Bill Gates" et "Microsoft" sont répertoriés comme un triplet (entité1 : Bill Gates, entité : Microsoft, relation : fondateur\_de) dans Freebase, alors ces entités représentent la relation "fondateur\_de".

#### 2.3.4.2.5 Approche basé apprentissage profond

L'utilisation de l'apprentissage profond dans l'extraction de relations a considérablement révolutionné la façon dont nous abordons cette tâche. Les réseaux de neurones profonds permettent une représentation complexe et hiérarchique des données textuelles, ce qui les rend particulièrement adaptés à la capture de nuances et de contextes subtils présents dans les relations entre entités. Par exemple, les architectures telles que les réseaux de neurones récurrents (RNN) ou les transformers peuvent prendre en compte la séquentialité du langage naturel, permettant ainsi une compréhension plus profonde des dépendances contextuelles. De plus, les modèles de langue pré-entraînés, comme BERT, ont démontré leur capacité à capturer des informations sémantiques riches, ce qui est crucial pour l'extraction précise des relations.

### 2.3.5 Similarité des phrases

### 2.3.5.1 Introduction

La similarité des phrases est la tâche qui consiste à déterminer à quel point deux textes sont similaires. Les modèles de similarité des phrases convertissent les textes d'entrée en vecteurs (embeddings) qui capturent les informations sémantiques et calculent à quel point ils sont proches (similaires) entre eux. Cette tâche est particulièrement utile pour la recherche d'informations et le clustering.[[17]](#footnote-17)

### 2.3.5.2 Approches pour la similarité des phrases

Plusieurs méthodes existent :

#### 2.3.5.2.1 Approches basé vecteurs sparses :

Ces approches utilisant les représentations discutées précédemment, Bag of words et tf-idf.

L'utilisation des représentations en sac de mots (Bag of Words) et du TF-IDF (Term Frequency-Inverse Document Frequency) pour mesurer la similarité des phrases est une méthode couramment employée en traitement automatique des langues. Le modèle sac de mots transforme un texte en une matrice de termes, où chaque phrase est représentée par la fréquence des mots qu'elle contient. Bien que simple, cette approche ignore l'ordre des mots, ce qui peut limiter sa capacité à saisir le sens contextuel. Pour améliorer cette représentation, le TF-IDF pondère les fréquences des termes en tenant compte de leur importance dans le corpus, réduisant l'impact des mots courants et accentuant ceux plus significatifs. En utilisant ces représentations, les phrases peuvent être comparées à l'aide de mesures de similarité telles que le cosinus, permettant d'évaluer la proximité sémantique entre elles. Bien que ces méthodes soient moins sophistiquées que les modèles d'apprentissage profond modernes, elles restent efficaces et largement utilisées pour des tâches telles que la recherche d'informations, le clustering de textes et la classification de documents.

L'utilisation des représentations en sac de mots (Bag of Words) et du TF-IDF pour mesurer la similarité des phrases présente plusieurs inconvénients, notamment la création de vecteurs clairsemés (sparse). Ces méthodes transforment chaque phrase en un vecteur de grande dimension où chaque dimension correspond à un mot unique du corpus. Étant donné que chaque phrase ne contient qu'un sous-ensemble des mots possibles, la plupart des dimensions de ces vecteurs sont nulles, ce qui conduit à des vecteurs extrêmement clairsemés.

Cette sparsité pose plusieurs problèmes. Premièrement, elle rend difficile la capture des relations sémantiques entre les mots, car les représentations ne prennent pas en compte l'ordre des mots ni les contextes dans lesquels ils apparaissent. Deuxièmement, la haute dimensionnalité des vecteurs peut entraîner des inefficacités en termes de stockage et de calcul, rendant les opérations de similarité, comme la mesure de la similarité cosinus, plus coûteuses et moins robustes. Enfin, les vecteurs clairsemés sont sensibles au bruit et peuvent manquer de généralisation, ce qui limite leur efficacité pour des tâches complexes nécessitant une compréhension fine des nuances linguistiques. Ces limitations ont conduit à l'adoption de méthodes plus avancées, telles que les embeddings de mots et les modèles de langue pré-entraînés, qui offrent des représentations denses et contextuelles mieux adaptées à la tâche de similarité de phrases.

#### 2.3.5.2.1 Approches basé embeddings :

L'utilisation des embeddings a transformé la manière dont nous abordons la similarité des phrases en offrant des représentations vectorielles denses et contextuelles des mots et des documents. Les word embeddings, tels que Word2Vec et GloVe, capturent des relations sémantiques entre les mots en les plaçant dans un espace vectoriel continu où des mots similaires sont proches les uns des autres. Ces représentations surpassent les modèles traditionnels comme le sac de mots et TF-IDF en considérant le contexte dans lequel les mots apparaissent. Cependant, les word embeddings présentent des limitations. Ils ne capturent que les relations de mots isolés et ignorent les nuances contextuelles des phrases. De plus, les embeddings statiques ne changent pas en fonction du contexte, ce qui peut conduire à des ambiguïtés pour des mots ayant plusieurs sens.

Pour pallier ces problèmes, des méthodes comme Doc2Vec ont été développées pour générer des embeddings de documents en prenant en compte l'ensemble de la phrase ou du texte. Doc2Vec produit des vecteurs denses pour des phrases, des paragraphes ou des documents entiers, offrant ainsi une meilleure capture des contextes larges et des relations sémantiques plus globales. Néanmoins, Doc2Vec peut être complexe à former et nécessite une quantité substantielle de données pour produire des représentations de haute qualité.

Des avancées plus récentes, telles que les modèles de langage contextuels comme BERT et GPT, utilisent des techniques de pré-entraînement sur de vastes corpus textuels pour créer des embeddings qui changent en fonction du contexte. Ces modèles offrent des représentations contextuelles riches, capturant les nuances fines des phrases et permettant une compréhension approfondie du langage naturel. Malgré leurs performances impressionnantes, ces modèles nécessitent une grande puissance de calcul pour l'entraînement et l'inférence, ce qui peut être un obstacle pour certaines applications.

## 2.4 **Travaux Connexes**

Dans le domaine de l'analyse des entretiens des forces de l'ordre, en particulier en ce qui concerne la détection des contradictions et la recommandation de questions, il existe une absence notable de systèmes ou de projets existants directement comparables à la portée et aux objectifs de notre travail. Alors que les technologies de l'intelligence artificielle (IA) ont été largement adoptées dans divers secteurs et applications, leur intégration dans le contexte spécialisé de l'analyse des entretiens des forces de l'ordre reste relativement peu explorée.

Malgré la reconnaissance croissante du potentiel de l'IA pour améliorer les processus d'investigation, notamment dans des tâches telles que l'examen de l'écriture manuscrite[[18]](#footnote-18), la littérature et les systèmes existants ignorent largement les exigences nuancées de l'analyse des entretiens des forces de l'ordre. L'absence de systèmes dédiés à la détection des contradictions et à la recommandation de questions dans ce domaine souligne la nécessité d'approches innovantes adaptées aux défis spécifiques et aux objectifs des agences d'investigation.

De plus, bien que les systèmes pilotés par l'IA pour l'analyse textuelle et les applications de traitement automatique du langage naturel (TALN) aient été largement étudiés dans d'autres domaines, tels que le service client ou les soins de santé, leur adaptation aux subtilités de l'analyse des entretiens des forces de l'ordre est limitée. Les recherches et projets existants dans des domaines adjacents ne parviennent pas à répondre aux exigences uniques de l'analyse des entretiens dans les contextes d'enquête, laissant un écart significatif dans les capacités requises pour des pratiques d'investigation complètes et efficaces.

## 2.5 Recherches dans la Contradiction

La détection de contradiction en traitement automatique du langage naturel (TALN) implique l'identification et le traitement des énoncés qui sont mutuellement exclusifs ou en conflit de signification. Cette tâche est cruciale dans diverses applications de TALN telles que l'analyse de sentiment, la réponse aux questions et les systèmes de dialogue, où la compréhension de la cohérence et de la consistance des entrées textuelles est essentielle.

Une approche courante de la détection de contradiction consiste à utiliser des modèles d'apprentissage automatique, en particulier des réseaux neuronaux. Ces modèles sont entraînés sur des ensembles de données annotés contenant des paires de phrases étiquetées comme contradictoires ou non-contradictoires[[19]](#footnote-19). Des techniques telles que les réseaux siamois ou les architectures basées sur les transformateurs comme BERT (Bidirectional Encoder Representations from Transformers) sont largement utilisées à cette fin. En apprenant des représentations contextuelles des phrases, ces modèles peuvent efficacement capturer les relations sémantiques et identifier les contradictions.

Une autre technique consiste à exploiter les caractéristiques linguistiques et les règles d'inférence logique pour détecter les contradictions. Cette approche implique souvent de représenter les phrases sous une forme structurée, telle que des formes logiques ou des graphes sémantiques, puis d'appliquer des techniques de raisonnement formel pour vérifier les incohérences. Cependant, cette approche peut rencontrer des défis pour gérer l'ambiguïté et la scalabilité.

De plus, des recherches récentes ont exploré l'intégration de sources de connaissances externes, telles que les bases de connaissances du bon sens ou les graphes de connaissances mondiaux, pour améliorer la détection de contradiction. En incorporant des connaissances de fond sur le monde, les modèles peuvent mieux comprendre les indices contextuels et formuler des jugements plus éclairés sur la plausibilité des énoncés.

## 2.6 Systèmes de recommandation

Les systèmes de recommandation sont des outils essentiels dans divers domaines, allant du commerce en ligne aux plateformes de streaming, en passant par les réseaux sociaux. Ils utilisent des techniques d'apprentissage automatique et de traitement du langage naturel pour analyser les préférences et les comportements des utilisateurs afin de proposer des contenus ou des produits pertinents. Par exemple, Amazon et Netflix emploient des algorithmes de filtrage collaboratif et de filtrage basé sur le contenu pour améliorer l'expérience utilisateur[[20]](#footnote-20). En combinant des méthodes telles que les réseaux de neurones et les modèles de graphes de connaissances, ces systèmes peuvent fournir des recommandations personnalisées et contextuellement adaptées [[21]](#footnote-21). De plus, les avancées récentes dans l'utilisation des modèles de transformateurs, comme BERT et GPT, ont permis d'améliorer considérablement la précision des recommandations en comprenant mieux le contexte et les intentions des utilisateurs [[22]](#footnote-22).

## 2.7 Recommandation des questions

La recherche spécifique dédiée uniquement aux systèmes de recommandation de questions dans l'analyse des entretiens des forces de l'ordre est limitée, mais le domaine le plus large des systèmes de recommandation et de recherche d'informations fournit des méthodologies pertinentes applicables à ce domaine. Les systèmes de recommandation de questions visent à aider le personnel de l'application de la loi à formuler des questions pertinentes lors d'enquêtes ou d'interrogatoires. Les techniques de traitement automatique du langage naturel (TALN) et d'apprentissage automatique sont souvent exploitées pour analyser les données textuelles et suggérer des questions pertinentes en fonction du contexte, des données historiques et des connaissances du domaine[[23]](#footnote-23). Ces systèmes peuvent utiliser des approches telles que le filtrage collaboratif, le filtrage basé sur le contenu ou des méthodes hybrides pour générer des recommandations de questions personnalisées adaptées à des scénarios d'enquête spécifiques.

## 2.8 Filtrage basé sur le contenu (Content-based filtering)

Le filtrage basé sur le contenu est une technique clé dans les systèmes de recommandation, utilisée pour proposer des éléments similaires à ceux que l'utilisateur a déjà appréciés. Contrairement au filtrage collaboratif, qui repose sur les préférences des autres utilisateurs, le filtrage basé sur le contenu analyse les caractéristiques des éléments eux-mêmes pour fournir des recommandations. Par exemple, dans le domaine de la musique ou des films, les systèmes de recommandation peuvent utiliser des informations telles que les genres, les acteurs, ou les artistes pour suggérer des contenus similaires [[24]](#footnote-24). Cette méthode utilise souvent des techniques de traitement du langage naturel pour extraire et analyser les caractéristiques des textes, telles que les descriptions de produits ou les résumés de films [[25]](#footnote-25). De plus, l'intégration des réseaux de neurones et des modèles de représentation de texte comme TF-IDF ou les embeddings de mots a permis d'améliorer la précision et la pertinence des recommandations basées sur le contenu [[26]](#footnote-26).

# Chapitre 3 : Conception

## 3.1 Introduction

Dans ce chapitre, nous détaillerons la conception de notre système. Nous commençons par présenter les diagrammes de cas d’utilisations et de classes pour avoir une vue globale du système. Puis nous passons à nos modules de détection de la contradiction et de recommandation des questions.

## 3.2 Diagramme de cas d’utilisation

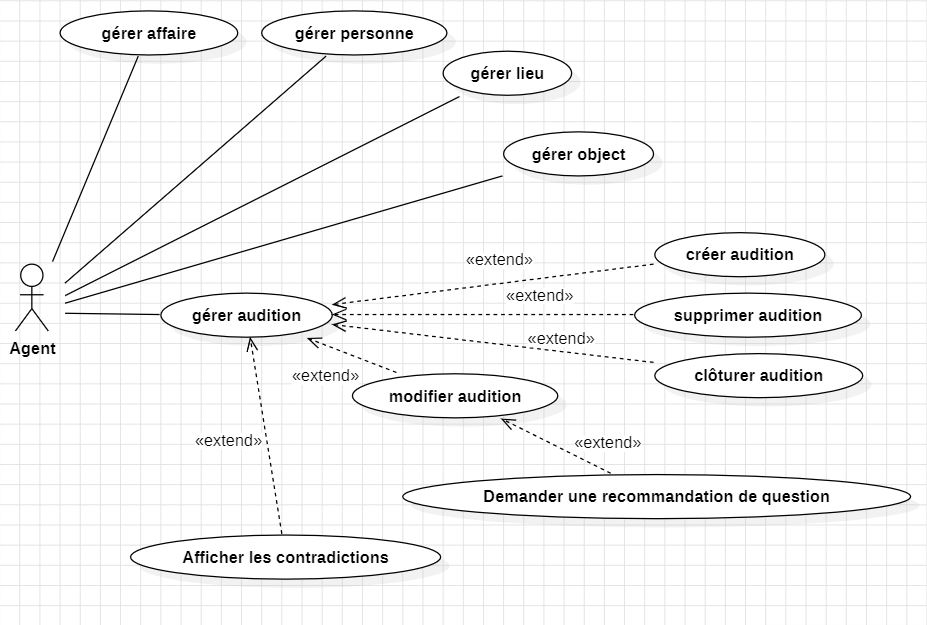


Figure 8 - diagramme de cas d'utilisation

## 3.3 Diagramme de classes

To be done

## 3.4 Module de détection de la contradiction

Notre approche consiste à proposer un modèle de reconnaissance d’entités nommées, puis combiner ce dernier avec un modèle d’extraction de relation afin d’avoir les relations définies pour ensuite détecter les contradictions. Ce processus est exécuté à chaque entrée d’une paire question-réponse.

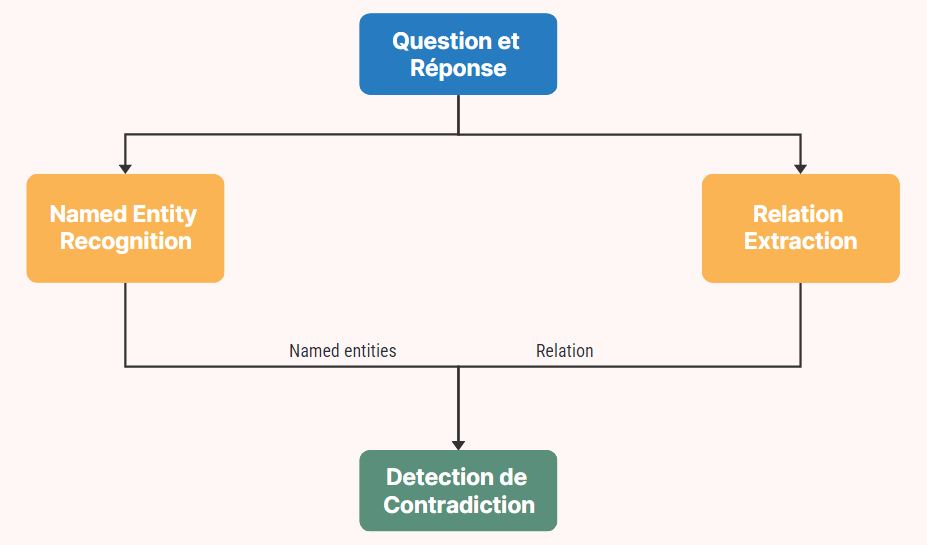


Figure 9 - Schéma global de la détection de la contradiction

## 3.5 Module de recommandation des questions

### 3.5.1 Introduction

Le module de recommandation des questions vise à assister les agents de la force publique pendant les interviews en générant des questions pertinentes et contextuelles. Ce module est crucial pour maintenir le flux de l'entretien, surtout lorsque l'agent ne sait plus quelles questions poser pour obtenir des informations complètes et cohérentes de la part de l'interviewé.

### 3.5.2 Description de l’approche

Pour la Recommandation de questions, on fait un content-based filtering pour choisir les auditions qui ont des features communes avec l’audition en question. Puis, pour chaque audition résultante, on passe leurs paires question-réponse au modèle de similarité de phrases pour trouver les paires les plus similaires à la dernière paire de l’audition en question, ce qui nous donne les questions suivantes à recommander.

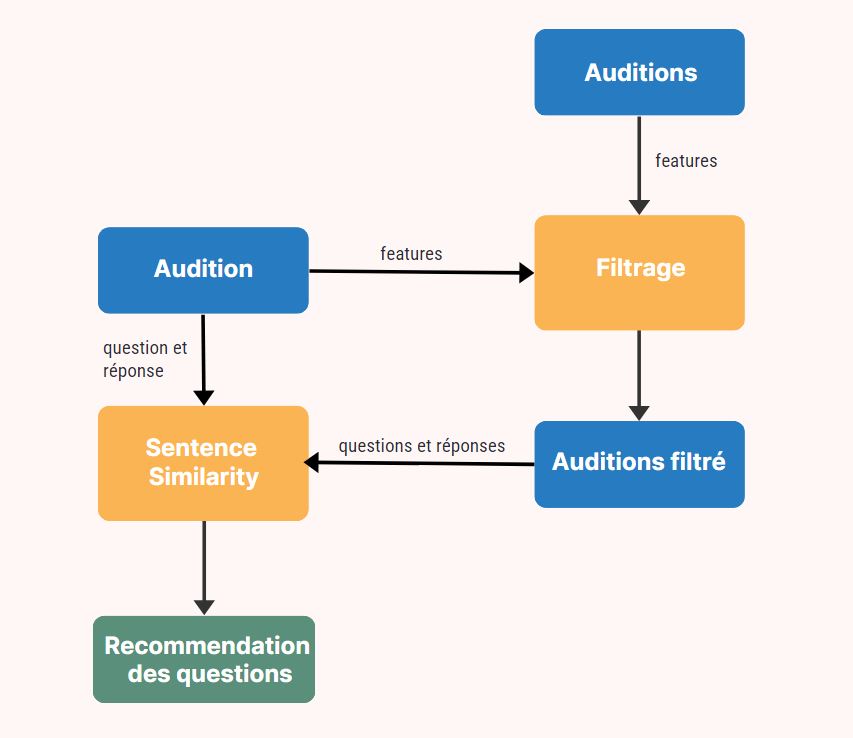


Figure 10 - Schéma global de la recommendation des questions

### 3.5.3 Filtrage

### 3.5.4 Sentence similarity

Après avoir identifié les auditions pertinentes, nous procéderons à une comparaison entre la dernière paire question-réponse de notre audition courante et les paires contenues dans ces auditions pertinentes. Pour ce faire, nous utiliserons la similarité de texte.

La similarité de phrase se réfère au degré de similarité ou de proximité entre deux phrases en termes de leur signification ou contenu sémantique. La similarité retournée est un score entre 0 et 1, 0 pour pas de relation entre les phrases, et 1 pour deux phrases identiques.

Il existe plusieurs approches pour faire la similarité de textes.

1. https://thideai.com/ai-in-law-enforcement-predictive-policing-and-crime-analysis/ [↑](#footnote-ref-1)
2. https://www.police1.com/iacp/articles/advancing-policing-through-ai-insights-from-the-global-law-enforcement-community-3SzYuRViccy8vwQ3/ [↑](#footnote-ref-2)
3. https://www.mdn.dz/site\_cgn/sommaire/presentation/unit\_spe/incc/incc\_fr.php [↑](#footnote-ref-3)
4. LGoldberg, Y. (2016). "A Primer on Neural Network Models for Natural Language Processing." Journal of Artificial Intelligence Research, 57, 345-420. [↑](#footnote-ref-4)
5. Jurafsky, D., & Martin, J. H. (2019). "Speech and Language Processing." Pearson. [↑](#footnote-ref-5)
6. Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. [↑](#footnote-ref-6)
7. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and WalterDaelemans, editors, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1532–1543. ACL, 2014. [↑](#footnote-ref-7)
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

   Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.

   In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob

   Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neu-

   ral Information Processing Systems 30: Annual Conference on Neural Informa-

   tion Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages

   5998–6008, 2017. [↑](#footnote-ref-8)
9. acob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT:

   Pre-training of deep bidirectional transformers for language understanding. In

   Proceedings of the 2019 Conference of the North American Chapter of the Asso-

   ciation for Computational Linguistics: Human Language Technologies, Volume 1

   (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019.

   Association for Computational Linguistics. [↑](#footnote-ref-9)
10. Hridoy Jyoti Mahanta. A STUDY ON THE APPROACHES OF DEVELOPING

    a NAMED ENTITY RECOGNITION TOOL. 02(14):58–61. [↑](#footnote-ref-10)
11. Pang, Bo, and Lee, Lillian 2008. Opinion mining and sentiment analysis. Foundations and Trends R⃝ in Information Retrieval, 2(1–2), 1–135. [↑](#footnote-ref-11)
12. Socher, Richard, Perelygin, Alex, Wu, Jean, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. Pages 1631–1642 of Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. [↑](#footnote-ref-12)
13. Maas, Andrew L., Daly, Raymond E., Pham, Peter T., et al. 2011. Learning word vectors for sentiment analysis. Pages 142–150 of Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, OR: Association for Computational Linguistics. [↑](#footnote-ref-13)
14. Krallinger, Martin, Leitner, Florian, Rodriguez-Penagos, Carlos, and Valencia, Alfonso. 2008. Overview of the protein–protein interaction annotation extraction task of BioCreative II. Genome Biology, 9(2), 1–19 [↑](#footnote-ref-14)
15. Sundheim, Beth M. 1992. Overview of the fourth message understanding evaluation

    and conference. Technical report. Naval Command Control and Ocean Surveillance Center, RDT & E Division, San Diego, CA. [↑](#footnote-ref-15)
16. Kartik Detroja, C.K. Bhensdadia, Brijesh S. Bhatt, A survey on Relation Extraction, Intelligent Systems with Applications, Volume 19, 2023, 200244, ISSN 2667-3053. [↑](#footnote-ref-16)
17. https://huggingface.co/tasks/sentence-similarity [↑](#footnote-ref-17)
18. Yavorsky, M. А., Useev, R. Z., & Kurushin, S. A. (2021). Information Technologies in Law Enforcement: Overview of Implements and Opportunities. *European Proceedings of Social and Behavioural Sciences (EpSBS)*, 4(2), 166. [↑](#footnote-ref-18)
19. Bowman, Samuel R., et al. "A large annotated corpus for learning natural language inference." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015. [↑](#footnote-ref-19)
20. Linden, Greg, Brent Smith, and Jeremy York. "Amazon.com recommendations: Item-to-item collaborative filtering." *IEEE Internet Computing* 7.1 (2003) : 76-80. [↑](#footnote-ref-20)
21. Shi, Yue, et al. "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges." *ACM Computing Surveys (CSUR)* 47.1 (2014): 3. [↑](#footnote-ref-21)
22. Sun, Yu, et al. "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformers." *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019. [↑](#footnote-ref-22)
23. Hu, Zhengyu, et al. "Personalized question recommendation in community question answering websites." ACM Transactions on Information Systems (TOIS) 33.2 (2015): 7. [↑](#footnote-ref-23)
24. Lops, Pasquale, Marco de Gemmis, and Giovanni Semeraro. "Content-based recommender systems: State of the art and trends." *Recommender systems handbook*. Springer, Boston, MA, 2011. 73-105. [↑](#footnote-ref-24)
25. Pazzani, Michael J., and Daniel Billsus. "Content-based recommendation systems." *The adaptive web*. Springer, Berlin, Heidelberg, 2007. 325-341. [↑](#footnote-ref-25)
26. Karatzoglou, Alexandros, et al. "Content-based recommendation systems." *Recommender Systems Handbook*. Springer, 2015. 627-681. [↑](#footnote-ref-26)