Université de BLIDA 1

Faculté des Sciences

Département d'Informatique

MEMOIRE DE MASTER

Option : Traitement Automatique de la Langue

SYSTEME D'EXTRACTION DES EVENEMENT ET DES EXPRESSION TEMPORELLES DES ARTICLES ARABE

Etabli par :

ALLACHE Fayçal

TAHRAOUI Abdelkader

Devant un jury composé de :

Mme TOUBALINE Nesrin	Présidente	USDB 1
Mme NASERI Ahlem	Examinatrice	USDB 1
M. ABBACHE Ahmed	Encadreur	UHBC
M. CHERIF-Zahar Amine	Promoteur	USDB 1

Année Universitaire 2018/2019

ملخص

الغرض من هذا العمل هو اختيار طريقة لاستخراج الأحداث والتعبيرات الزمنية من مقالات اللغة العربية. ثم تحسينها وتطبيقها، لتقييمها أخيرًا على مجموعة من المقالات من فئات مختلفة.

يستند النهج على قواعد مبرمجة في Python، حيث تم تحسين هذه الأخيرة لكي تستخرج العديد من الأحداث، وكذلك على برامج ألية (AUTOMATES) لقواعد تم إنشاؤها في Unitex، لكي تشمل عدة أنواع وأشكال من التعبيرات الزمنية.

الكلمات المفتاحية

العربية، التعبير الزمني، استخراج الحدث ، القواعد، معالجة الآلية اللغة العربية،Python، Unitex

Résumé

L'objet de ce présent travail est de choisir une approche d'extraction des évènements et des expressions temporelles sur des articles en langue arabe standard. Puis l'améliorer, l'implémenter pour enfin l'évaluer sur un corpus d'articles de différentes catégories.

L'approche en question est celle à base de règles et de grammaire programmée en Python, où les règles ont été améliorées de telle sorte à ce qu'elles renvoient plusieurs évènements, ainsi que des grammaires crées en Unitex touchant différents types et formes d'expressions temporelles.

Mots clés :

Arabe, expression temporelles, Extraction événement, grammaire, python, traitements automatique de la langue arabe, TALA, Unitex.

Abstract

The purpose of this work is to choose an approach for extracting events and temporal expressions from standard Arabic language articles. Then improve it, implement it, finally evaluate it on a corpus of articles of different categories.

The approach is based on rules and grammars programmed in Python, where the rules have been improved so that they return multiple events, as well as grammars created in Unitex that affect different types and forms of temporal expressions. .

KeyWords

Arabic, Event Extraction, Arabic Naturel Language Processing, grammar, python, Temporal Expression, Unitex.

Remerciement

Tout d’abord, nous rendons grâce à Allah Tout-Puissant de nous avoir doté du savoir, de la santé et de la volonté.

Nous tenons à remercier plus particulièrement Mr.Abbache Ahmed pour ses efforts, grâce auxquels le déroulement de nos projets se fut dans les meilleures conditions.

Nous adressons également nos remerciements à notre promoteur Mr. CHERIF-Zahar Amine.

Nous remercions également Mme Benblidia, Mme Oukid Saliha, Mme Mezzi Melyara, Mme Ouahrani Leila, Mme Arkam Malek.

Un grand merci particulier pour le Chef du Département de l’informatique et Mme Ali Masoude vice Doyenne de la Faculté du Sciences qui n’a ménagé aucun effort pour le bon déroulement de notre cursus.

Nous tenons à exprimer nos sincères remerciements à tous les enseignements de la qualité des cours dispensés par le Master « TAL » avoir excité nos réflexions .et une profonde satisfaction intellectuelle, merci donc aux enseignants-chercheurs.

Nous tenons à présenter notre reconnaissance aux membres du jury pour nous avoir fait l’honneur d’accepter d’évaluer ce travail.

Nos remerciements vont également à toute personne ayant contribué de près ou de loin à la concrétisation de ce travail.

Merci à toutes et tous.

Sommaire

Liste des Figures

Liste des tableaux

INTRODUCTION

L'avancement technologique des moteurs de recherche et d'internet en général a induit une accumulation fulgurante des informations textuelles électroniques. Il est alors indispensable de se munir d'outils de fouille afin de mieux traiter et évaluer la pertinence de ces dernières.

L'extraction des expressions temporelles et des évènements consiste à identifier les contenus d'informations dans un texte afin de faciliter sa classification. Elle a inspiré de diverses orientations sous différentes approches, on site :

- Les approches basées sur des règles se concentrant sur l'extraction en utilisant de nombreux jeux de règles créées ;
- Les approches basées sur des algorithmes d'apprentissage automatiques ;
- Les approches hybrides qui englobent les deux types précédents dans une même approche.

De nos jours, la plupart des systèmes d'extraction des expressions temporelles et des évènements traitent des textes en langues indo-européennes (l'anglais, le français, etc.). Le besoin de développer des systèmes d'extraction des expressions temporelles et des évènements dédiés pour la langue arabe devient de plus en plus incontournable. En effet, nous avons recensés uniquement six travaux :

- Pour l'extraction des évènements : Approche pour l'extraction d'évènement TF-IDF et cooccurrence **Amina Chouigui, Oussama Ben Khiroun1, et Bilel Elayeb** [1]. l'approche basée sur la connaissance **Mohamed [2]**. L'approche hybrides **Anup Kumar Kolya, Asif Ekbal, et Sivaji Bandyopadhyay** [3].
- Pour l'extraction des expressions temporelles : une approche d'apprentissage automatique **Iman Saleh, Lamia Tounsi, et Josef van Genabith** [4], une approche basée sur des règles de **aliane** [5] et le travail de **Khaled Shaalan et Hafsa Raza** [6].

De plus, les approches sur la langue arabe trouvées en littérature sont rares et sont implémentées pour être appliquées sur des tweets. Malheureusement aucune approche n'est applicable sur des articles.

L'objet de notre modeste mémoire est d'étudier, de proposer et d'implémenter une approche basée sur des règles (grammaires) que nous appliquerons à des articles en langue arabe de différentes catégories. Notre mémoire est divisé en quatre chapitres :

- Une première chapitre ou on définira les différents termes liés à l'extraction de données en général.
- Dans le deuxième chapitre, nous tenterons de cerner l'objet d'étude, à savoir l'évènement et l'éventail d'approches existantes pour sa génération d'une façon automatique. Nous introduirions alors la terminologie nécessaire et nous examinerons le processus d'extraction des évènements. Par la suite, nous donnerons une illustration assez exhaustive des approches existantes pour la génération d'un extrait.
- Le troisième chapitre est dédié à la conception et à l'approche proposée
- On verra dans le quatrième chapitre l'implémentation et l'évaluation du programme.

PARTIE 1 : ETAT DE L'ART

Chapitre 1. Extraction d'informations

1.1 Introduction

Une entreprise veut suivre les sentiments généraux à propos de sa nouvelle version produit dans les blogs Web. Une autre entreprise veut utiliser les flux de nouvelles acheté à une agence de presse pour construire un aperçu détaillé de toutes les technologies les tendances dans le développement des technologies des semi-conducteurs. La société veut également un calendrier de toutes les transactions commerciales impliquées dans ce développement. Une agence spatiale permet aux astronautes d'interroger de grandes quantités de documentation technique au moyen du langage naturel. Un gouvernement recueille des données sur une catastrophe naturelle et souhaite informer de toute urgence services d'urgence avec un résumé des dernières données disponibles. Une agence intelligente enquête sur les tendances générales des activités terroristes partout dans le monde. Ils ont une base de données de millions de flux de nouvelles, minutes et e-mails et souhaitez les utiliser pour obtenir un aperçu détaillé de tous les terroristes. Événements survenus dans une région géographique donnée au cours des cinq dernières années. Il y a des dizaines de milliers d'articles, de documents de conférence et rapports techniques à étudier.

L'extraction d'informations est généralement associée à des modèles extraction des informations sur les événements à partir de texte en langage naturel, ce qui était une tâche populaire de la Message Understanding Conferencesde la fin des années quatre-vingt et des années quatre-vingt-dix (Sundheim, 1992). MUC tâches d'extraction d'informations à partir d'un ensemble prédéfini de modèles, chacun contenant des informations spécifiques slots qui codent des types d'événements pertinents pour un sujet très spécifique domaine - par exemple le terrorisme en Amérique latine - et a utilisé relativement techniques simples d'appariement de modèles pour remplir ces modèles avec des exemples spécifiques de ces événements à partir d'un corpus de textes. Modèles dans la forme de grammaire ou de règles (par exemple, sous la forme d'expressions régulières) étaient mappé sur le texte afin d'identifier l'information.

MUC a été le premier effort à grande échelle visant à stimuler la recherche dans le domaine d'extraction automatique d'information et il définirait le domaine de la recherche pour les

décennies venir. Même au moment de l'écriture, l'extraction d'informations est souvent associée avec des techniques de correspondance de modèle basées sur des modèles. Sans surprise, l'héritage de la CUM résonne encore très fortement dans la définition de Riloff et Lorenzen d'extraction d'informations :

> Les systèmes IE extraient des informations spécifiques à un domaine du texte
> En langage naturel. Le domaine et les types d'informations à extraire doivent
> Être définis à l'avance. Les systèmes extraient informations se concentrent
> Souvent sur l'identification d'objet, telle que les références à des personnes,
> Des lieux, des sociétés et des objets physiques. [...] Des modèles d'extraction
> Spécifiques à un domaine (ou quelque chose de similaire) sont utilisés pour
> Identifier les information pertinences. [7]

Cette définition représente un point de vue traditionnel sur ce que l'extraction d'informations est et il capture plus ou moins de quoi traite cette discipline : L'extraction d'informations sémantiquement définies à partir d'un texte, à l'aide d'un ensemble d'extraction des règles adaptées à un domaine très spécifique. Les points principaux de cette définition est qu'un système d'extraction d'informations identifie l'information contenues dans un texte, c'est-à-dire dans une source d'informations non structurée, et les informations qui adhèrent à la sémantique prédéfinie (par exemple, des personnes, des lieux) etc.). Cependant, nous verrons dans la suite.

1.2 Définition text mining

Le texte mining est défini comme "le processus de recherche utile ou des modèles intéressants, des directions, des tendances, ou règles des textes non structurés".

Plusieurs techniques ont été proposées pour l'exploration de texte, y compris les structures conceptuelles, extraction de règles d'association, arbres de décision et les méthodes d'induction de règles. En outre, les techniques de recherche d'informations (IR) sont largement utilisées (Baeza-Yates & Ribeiro-Neto 1999) pour les tâches telles que la correspondance de documents, le classement et la mise en cluster.

1.3 Définition de l'Extraction d'information

L'extraction d'information est une technologie récente, qui cherche à répondre à un besoin très ancien : acquérir de la connaissance à partir des textes. Cette nécessité s'est accrue ces vingt dernières années avec l'essor considérable de la masse des documents disponibles au format électronique (internet, courriels et documentation électroniques) qu'il faut gérer afin d'extraire ou de filtrer les informations utiles et pertinentes parmi toutes celles contenues dans ces documents.

Comme la recherche d'information, le résumé automatique ou les systèmes de questions-réponses (QA), l'extraction d'information a l'ambition de répondre à ce défi, d'où le développement de nombreuses applications destinées à des institutions, au monde des affaires et/ou de l'industrie [8].

Ces informations sont destinées à créer ou alimenter un entrepôt de données. La tâche d'extraction est réalisée grâce au remplissage des formulaires prédéfinis (Template).

Ces formulaires, dits formulaires d'extraction, sont définis dans le but de représenter la connaissance à rechercher par une structure déterminée a priori. Ils décrivent un ensemble d'entités, les relations entre celles-ci et les événements impliquant ces entités. Par exemple, un formulaire concernant des accidents de la route devra spécifier des champs comme « Lieu de l'accident », « Nombre de victimes », « Identité des victimes » ou encore « Cause de l'accident ».

Les informations extraites par un système d'Extraction d'Information peuvent être consultées par des utilisateurs humains (par exemple via la génération de rapports d'événements), alimenter une base de données afin d'être analysées plus tard (interrogation par requêtes ou fouille de données).

Bogota, 30 Aug 89 (Inravision Television Cadena 2) — Last nigth's terrorist target was the Antioqua liqueur plant. Four powerful rockets were going to explode very close to the tanks where 300,000 gallons of the socalledcatille crude, used to operate the boilers, is stored. The watchmen on duty reported that at 20:30 they saw a man and a woman leaving a small suitcase near the fence that surrounds the plant. The watchmen exchanged fire with the terrorists who fled leaving behind the explosive material that also included dynamite and grenade rocket launchers, metropolitan police personnel specializing in explosives, defused the rockets. Some 100 people were working inside the plant.

The damage the rockets would have caused had they been activated cannot be estimated because the caribe soda factory and the Guayabal residential area would have also been affected.

The Antioquia liqueur plant has received threats in the past and maximum security has always been practised in the area. Security was stepped up last night after the incident. The liqueur industry is the largest foreign exchange producer for the department.

Figure 1.1: (Corpus et Formulaire MUC) [8].

Le formulaire suivant est rempli manuellement par des experts à partir du texte ci-dessus et correspond aux informations qui doivent être trouvées par les systèmes participant à l'évaluation MUC-3.

Tableau I-1: Exemple d'un formulaire [8]

Date of Incident	29 august 1989
Type of Incident	attemptedbombing
Category of Incident	Terroristact
Perpetrator: ID of Indiv(s)	“man” “woman”
Perpetrator: ID of Org(s)	-
Perpetrator : Confidence	-
Physical Target : ID(s)	“Antioquia liqueur plant”
Physical Target : Number	1
Physical Target : Type(s)	commercial : “Antioquia liqueur plant”
Human Target : ID(s)	“people”
Human Target : Number	Plural
Human Target : Type(s)	civilian : “people”
Target : Foreign Nation(s)	-
Instrument : Type(s)	-
Location of Incident	Colombia : Antioquia (Department)
Effects on Physical Target(s)	no damage: “Antioquia liqueur plant”
Effects on Human Target(s)	no injury or death: “people”

Le signe « - » indique que le champ n’est pas renseigné, c’est-à-dire qu’aucune information n’a été trouvée pour le remplir.

1.4 ConférencesMUC – Message Understanding Conferences

Les recherches actuelles en EI ont été influencées par les conférences MUC (Grishman & Sundheim, 1996). Ces conférences qui se sont déroulées entre 1987 et 1998, faisaient partie du programme TIPSTER13 financé par DARPA14. Ce programme comportait trois tâches

- La détection des documents.
- L'extraction d'information.
- Le résumé de textes.

Les campagnes d'évaluation MUC ont été organisées afin de confronter les systèmes d'extraction d'information réalisés par différentes équipes en comparant leurs performances avec des mesures précises et objectives. Ces mesures, inspirées de celles définies pour le domaine de la recherche d'information, sont devenues un standard pour toute évaluation des résultats de l'EI. Ainsi, la précision mesure la qualité du système, c'est-à-dire le nombre d'informations extraites correctement par rapport au nombre d'informations extraites. Le rappel lui mesure la couverture du système, c'est-à-dire le nombre d'informations correctement extraites par rapport au nombre d'informations correctes présentes dans le corpus. Enfin, la F- mesure permet de disposer d'une évaluation globale du système en combinant précision et rappel. [9]

L'apport des conférences MUC a été considérable ; aussi bien en termes d'identification des problèmes à prendre en compte (linguistique, représentation des connaissances, acquisition de ressources, travail sur corpus…) qu'en termes de méthodes et de techniques pour les résoudre.

Les textes servant de support à l'évaluation provenaient de différents domaines. Les premières conférences ont porté sur l'extraction d'information à partir des messages militaires, par contre ce thème a été développé dans les conférences ultérieures pour couvrir les rapports de presse. Divers systèmes d'extraction d'information ont été testés sur différents types de textes : récits d'attentats (MUC-3 et MUC-4), annonces de produits (MUC-5), annonces financières concernant les prises de participation des entreprises (MUC-6), etc. Les systèmes en compétition devaient remplir un ou plusieurs formulaires fixés à l'avance en fonction du

domaine. Par exemple, pour les annonces financières, ils devaient extraire les différentes sociétés (acheteurs, vendeurs, achetées), la date, le lieu et le montant de la transaction financière, etc. [9].

Tableau I-2: Résumé des différents contenus de textes traités dans chaque conférence [9].

(MUC 1, 1987) et (MUC-2, 1989)	*Ont traité et analysé les rapports d'opérations tactiques navales.*
(MUC3, 1991) et (MUC4, 1992)	L'objectif était d'analyser des textes journalistiques traitant du terrorisme en Amérique Latine, afin d'extraire, des dépêches d'agence de presse, le maximum d'information sur des actes terroristes.
(MUC5, 1995)	Ont traité un corpus de nature économique pour extraire des informations de type fusion, rachat, et création d'entreprises internationales et la fabrication de circuits électroniques.
(MUC 6,1996)	Une suite de MUC 5, a traité les changements de dirigeants à la tête des entreprises.
(MUC7, 1998)	S'est intéressée à l'analyse de textes journalistiques rapportant des crashs d'avion et de tirs de missiles.

1.5 Les taches de l'extraction d'information

Les composants trouvés dans les systèmes d'EI d'aujourd'hui reflètent largement les tâches définies dans ces conférences. Les tâches de la dernière conférence, MUC-7, en 1998 (les plus difficiles dans la série) ont été les suivantes :

- Reconnaissance des entités nommées.
- Détection de la coréférence.
- Reconnaissance des éléments du formulaire.
- Reconnaissance des relations.
- Reconnaissance des scénarios (« scenario template »).

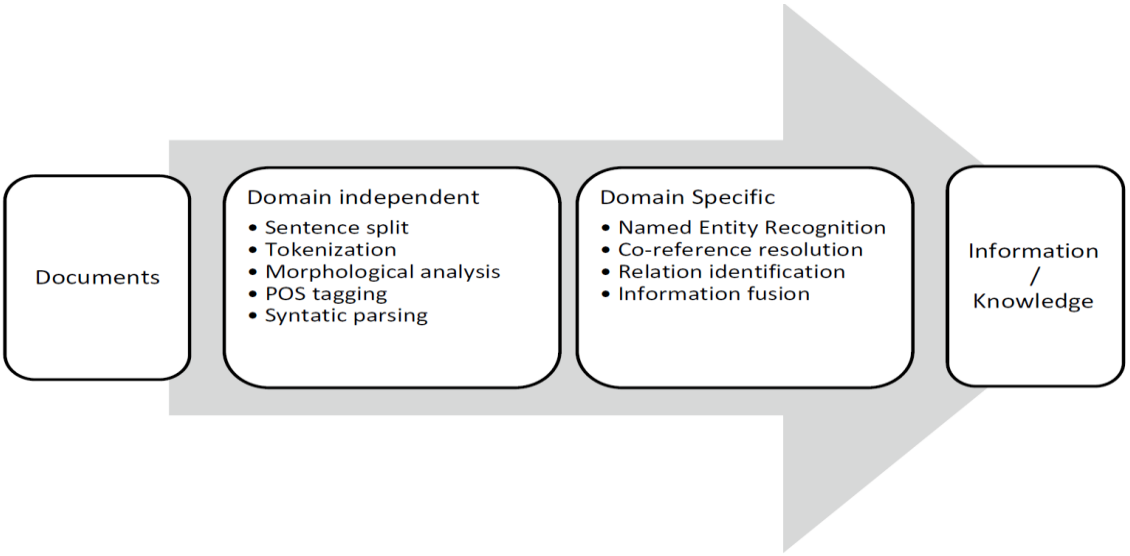

Figure 1.2: Traitement général des systèmes d'extraction d'information.

1.5.2 Reconnaissance des entités nommées

Cette tâche consiste à repérer toutes les formes linguistiques bien identifiées, à l'instar des noms propres de personnes, d'organisations, de lieux, etc. mais aussi les expressions temporelles (dates, durées…), les quantités (monétaires, unités de mesures, pourcentages…) et à leur affecter une étiquette sémantique choisie dans une liste prédéfinie [9].

1.5.3 Détection de la coréférence

Cette tâche consiste à repérer les groupes nominaux et les pronoms personnels co- référents et à les baliser dans les textes. Par exemple, dans « En 1963, Warda El-Djazairia épouse Djamel Kesri, un des fondateurs de l'ancienne sécurité militaire. Après son exil en Égypte, elle se rendait de moins en moins à son domicile à Alger. », La résolution des coréférences devrait relier « Elle » à « Warda El-Djazairia » [9].

1.5.4 Reconnaissance des éléments du formulaire

Cette tâche, qui repose sur les deux tâches précédentes, consiste à associer des informations (descriptions, informations complémentaires) aux entités reconnues. Elle associe en fait de l'information descriptive, généralement sous la forme de groupes nominaux, aux entités précédemment identifiées. Cette information descriptive correspond à un attribut de l'entité concernée [9].

1.5.5 Reconnaissance des relations

La reconnaissance des relations s'attache à identifier un certain nombre de relations, le plus souvent binaires, entre les entités extraites précédemment. Ainsi, dans l'exemple précédent, cette tâche permet de repérer une relation de mariage entre les entités personnes « Warda El-Djazairia » et « Djamel Kesri » [9].

1.5.6 Reconnaissance des Scenarios

Cette tâche relie entre les entités et les relations précédemment reconnues des descriptions d'évènement relatif au domaine étudié. Les différents traits complémentaires, telles que la localisation spatiale et temporelle sont également associés. La reconnaissance des scénarios est une tâche particulièrement difficile. Elle dépend des résultats des étapes précédentes et possède donc un score plus faible, dépendant de la composition de leurs résultats [9].

1.6 Les Mesures d'évaluation

Les résultats obtenus de système ou bien des méthodes doivent être non ambigüe pour toutes les chaines pertinentes dans le texte. La tâche nécessite que le système reconnaître ce que représente une chaîne, non seulement apparence superficielle. Malgré que l'apparence superficielle puisse être la bonne réponse si ce n'est pas le cas dans un nom qui peut représenter par un nom de personne, d'une localisation ou bien d'une organisation dans ce cas il faut utiliser des techniques qui tirent des informations d'un plus grand contexte ou à partir de listes de références. Deux mesures de notation on était développé le rappel et la précision ou :

Le rappel (R) : est une évaluation de la couverture du système. Il mesure la quantité de réponses pertinentes d'un système par rapport au nombre de réponses idéales

$$\mathbf{R} = \frac{\textbf{Nombred'entités correctes détectées}}{\textbf{Nombred'entités manuellement identifiées}} \textbf{(1)}$$

La précision (P) : est une évaluation du bruit du système. Elle mesure la proportion des réponses correctes parmi l'ensemble des réponses fournies par le système

$$P = \frac{\text{Nombred'entités correctes détectées}}{\text{Nombred'entités détectées}} (2)$$

Ces deux mesures de performance se combinent pour former une mesure de la performance F, qui est calculée par la moyenne harmonique pondérée uniformément de précision et rappel :

Le F-mesure (F) : C'est la moyenne harmonique de la précision et du rappel qui mesure la capacité du système. À donner toutes les solutions pertinentes et à refuser les autres, une mesure populaire qui combine la précision et le rappel est leur pondération.

$$F = \frac{2\,(P\,.\,R)}{(P+R)} (3)$$

Ces mesures peuvent être définies sur les formulaires ou sur un nombre limité de champs [Gri96]. A titre indicatif, pour la tâche MUC 7, la précision des systèmes est inférieure à 40% pour un rappel de 70%.

Dans MUC-7, une réponse correcte d'un chercheur de nom est une réponse dont l'étiquette et les deux limites sont correctes. Il existe trois types d'étiquettes, chacune utilisant un attribut pour spécifier une entité particulière. Les types d'étiquette et les entités qu'ils désignent sont définis comme suit :

(i) Entité (ENAMEX) : personne, organisation, lieu.

(ii) Expression temporelle (TIMEX) : date, heure.

(iii) Expression numérique (NUMEX) : argent, pourcentage.

Une réponse est à moitié correcte si l'étiquette (le type et l'attribut) est correcte mais qu'une seule limite est correcte.

Alternativement, une réponse est à moitié correcte si seulement le type de l'étiquette (et non l'attribut) et les deux limites sont correct [10].

1.7 Conclusion :

L’objet de ce chapitre est de définir les principaux termes utilisés dans ce mémoire, à travers les définitions et les étapes d’extraction des informations et leur développement au fil des années. Ainsi que les différentes mesures d’évaluation de ces dernières. Ce qui nous facilitera par la suite la décortication des approches.

Chapitre 2. Les approches d'extraction des évènements et des expressions temporelles

2.1 Introduction :

Les approches d'extraction des évènements et des expressions temporelles sont vraiment rares, surtout en ce qui concerne la langue arabe.

Les approches qui ont attirés notre attention sont ceux cités en introduction générale. On essayera dans ce qui suit, de les présenter afin de choisir une approche pour notre système.

2.2 Articles des extractions des expressions temporelles :

2.2.1 NERA: Named Entity Recognition for Arabic

Description

Dans cet article les auteurs ont développé un système de reconnaissance des entités nommées pour l'arabe (NERA) à l'aide d'une **approche à base de règles**. Les ressources créées été une liste représentant un dictionnaire de noms et une grammaire dans la forme d'expressions régulières, qui sont responsables de reconnaissant des entités nommées. Un mécanisme de filtrage est utilisé à deux fins différentes : (1) la révision des résultats de l'extracteur d'entité nommée et (2) la désambiguïsation de personnes identiques.

L'approche utilisée

Le système NERA nécessite deux principales ressources : une liste (répertoire toponymique) et une grammaire à états finis. Un mécanisme de filtrage est également utilisé qui active les capacités de révision dans le système. La Figure 2.1 montre l'architecture abstraite du système NERA. Le système convertit le texte Arabe d'entrée non structuré en forme structurée en produisant les annotations du EN Arabe à la suite de la tâche de reconnaissance.

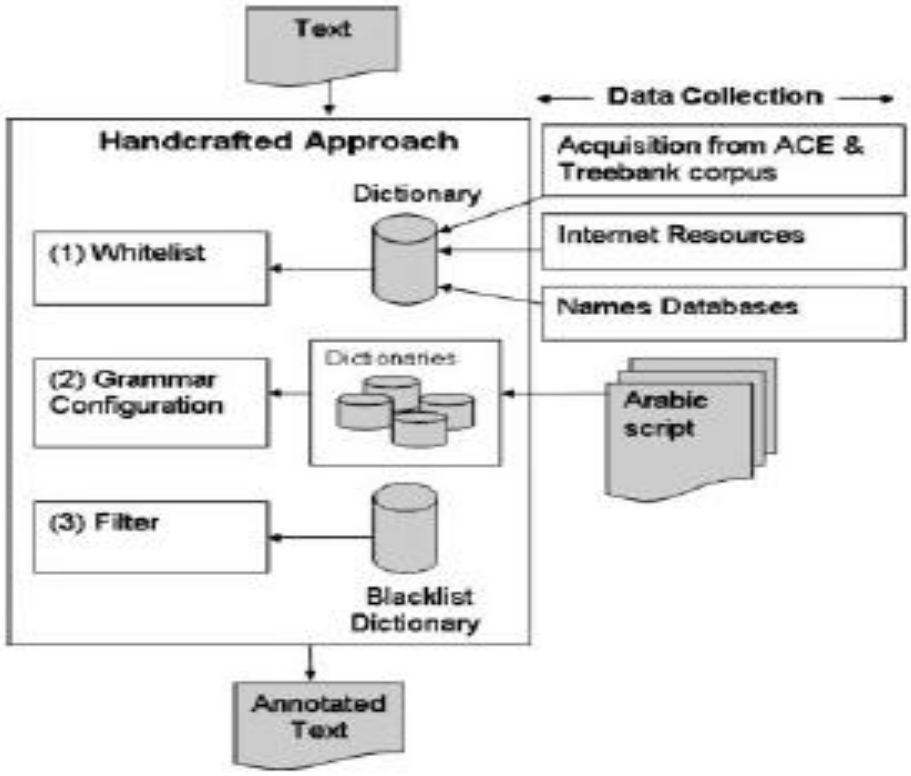

Figure 2.1: La technique de reconnaissance utilisée [6].

La technique de reconnaissance utilisée les trois étapes principales suivantes :

a. *La liste Blanche (whitelist) :* jouer le rôle d'un dictionnaire statique fixe de déférente ENs. Il contient des entrées dans le format :

الشيراوى عبدالرحمن قاسم|Abdulrahman Qasim Mohammed Alshirawi

b. *Grammaire :* La grammaire effectue la reconnaissance et l'extraction d'entités nommées arabe du texte d'entrée basé sur des règles dérivées. Il décrit les modèles pour faire correspondre les éléments de réseau, les annotations étant créées Par conséquent. En raison des particularités et de la complexité de la langue arabe, les règles de grammaire ce sont un essentiel traitement ressource pour le système de reconnaissance. En tant que langue agglutinante, l'arabe a de nombreuses formes infléchies. Par conséquent, les règles de grammaire construites codent d'informations morphologiques décrivant la structure infléchie des formes de mots candidats.

c. *Filtrage :* Un mécanisme de filtration est utilisé qui sert deux différentes fins : révision des résultats de l'extracteur EN et désambiguïsation des matchs retournés par différents extracteurs EN [6].

Discussion

Les résultats de l'évaluation semblent jusqu'à présent très prometteurs ; Le système NERA a été efficacement évalué sur un corpus étiqueté ; il a obtenu des résultats satisfaisants en termes de précision, de rappel et de F-mesure.

2.2.2 ZamAn and Raqm : Extraction d'expressions temporelles et numériques en arabe

Description

Dans cet article les auteurs ont développé le système ZamAn, qui utilise une méthode **d'apprentissage automatique,** développée pour étiqueter les expressions temporal Arabes.

L'approche utilisée

Ils ont utilisé Yamcha (Yet Another Multipurpose Chunk Annotator) pour détecter des expressions temporelles. Il utilise un algorithme d'apprentissage automatique base sur le Support Vector Machines (SVM).

La technique de reconnaissance utilisée les deux étapes principales suivantes :

a. *Prétraitement d'Arabic TreeBank (ATB) :* Le prétraitement comprend :
   i. **Variation orthographique** : ils réduisent toutes les variantes possibles du caractère Alif pour normaliser les données.
   ii. **Réduire l'étiquette** : L'étiquette ATB consiste en 492 étiquettes POS à grain fin, codant les caractéristiques morphologiques.
   iii. **Fractionnement de la ATB** : ils appliquent la scission d'arborescence habituelle (80% de formation, 10% de développement, 10% de test).
   iv. **Annotation manuelle pour –TMP**.

b. *Caractéristiques de classification* : Leur méthode est comparable aux travaux de Hacioglu[i] en ce qui concerne la méthode de détection et les fonctionnalités utilisées dans la détection. Ils ont utilisé 9 caractéristiques pour faire la classification [4].

Discussion

ZamAn est le premier système robuste et précis de reconnaissance et d'extraction des expressions temporelles de l'Arabe.

2.2.3 Annotation d'événements, expressions de l'heure et du lieu dans des textes arabes

Description

Dans cet article les auteurs ont développé une **approche non supervisée**, sans utiliser des corpus n'annoter ni de lexique (autrement dit sans utiliser une annotation POS et lexicon) mais ils ont utilisé un algorithme de segmentation non supervisé puis un ensemble de règles minimaliste permettant d'obtenir une annotation partielle POS du corpus. Ce dernier servira comme une base pour le processus de reconnaissance qui contient un ensemble de règles utilisant des marqueurs linguistiques spécifiques pour identifier les évènements et expressions temporelles.

L'approche utilisée

a. Détecter l'expression clé du texte arabe : ils ont proposé une approche sans prétraitement exhaustive comme le marquage POS et sans utiliser les dictionnaires, mais en utilisant une ontologie. Ensuite ils ont essayé d'identifier les évènements et les expressions temporelles en utilisant des indices de surface à partir du texte (sans entité nomes).
b. Segmentation et marquage POS partielle : ils ont voulu travailler en utilisant la surface indice des textes en adaptant l'algorithme (Aliane, 2011) qui est un algorithme de segmentation basé sur la théorie linguistique Arabe.
c. Ils ont construit manuellement des règles pour annoter les verbes et les noms dans les corpus en utilisant les affixes gauche et droite obtenus par le processus de segmentation et par une autre surface d'indice, d'où l'obtention de quatre règles.
d. Détection d'événement verbal : dans ce travail ils se sont intéressés par tout ce qui concerne l'annotation d'événements verbaux.
e. Détection d'expressions temporelles : Dans cette étape ils identifient des unités linguistiques non verbales qui véhiculent l'information temporelle en détectant

« marqueurs temporel », ensuite ils appliquent une analyse contextuelle à droite et à gauche des marqueurs identifiés.

f. Détection des expressions de lieu : afin de détecter les expressions de localisation, ils ont utilisé également la surface marqueurs des textes. Ces marqueurs sont autonomes marqueurs ou marqueurs de déclenchement [5].

Le schéma suivant regroupe et simplifie les différentes étapes du processus de l'approche

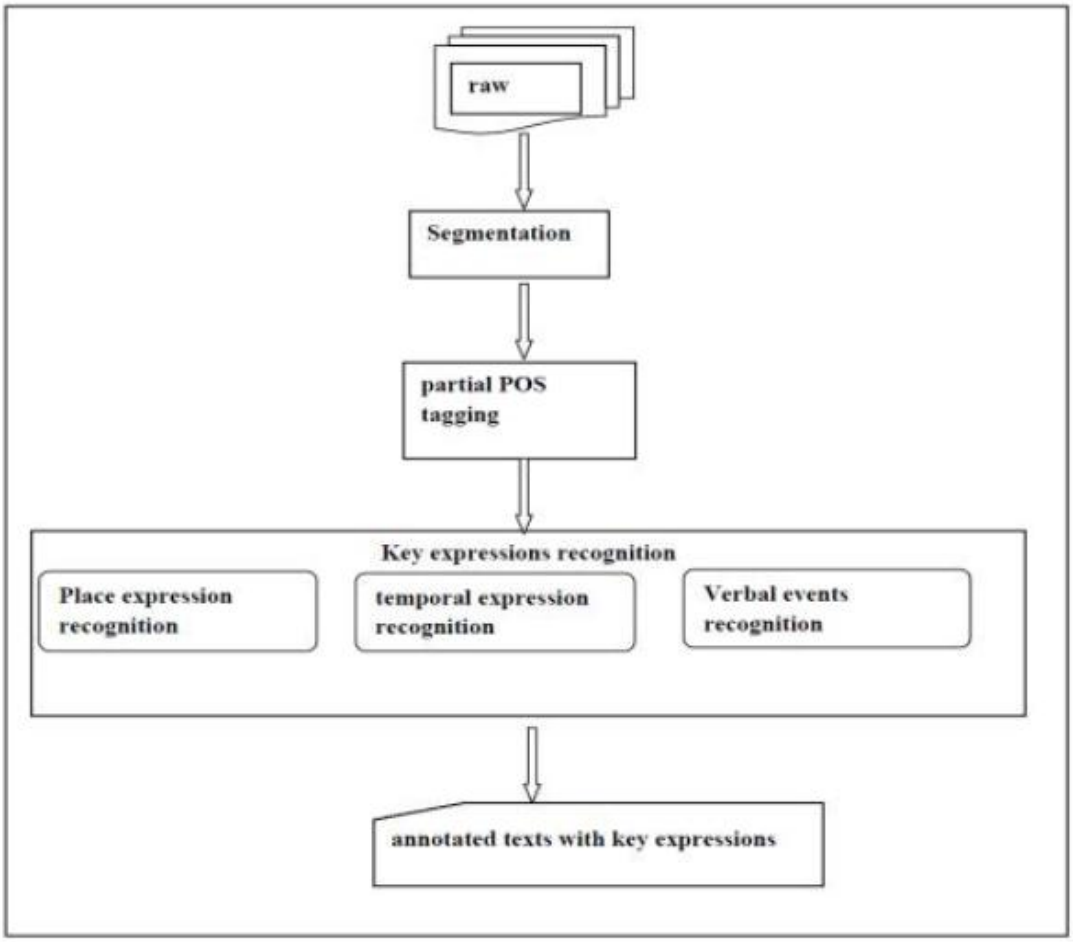

Figure 2.2: Architecture de l'approche [5].

Discussion

Les auteurs ont testé leur approche sur un corpus de 30 articles du web, écrit en Arabe standard moderne. Les textes ne sont pas voyelles. Le corpus annoté par les balises <événement> pour les événements verbaux détectés, <Timex> pour les expressions de temps et <Pl> pour les expressions de lieu. Le système était capable de reconnaître 168 événements verbaux sur 268 et

montre une F-mesure de 84% pour les expressions temporelles et 45% pour les expressions de lieu. Ces taux de reconnaissance sont influencés par les ambiguïtés laissées par l'étape de marquage partiel qui n'a pas détecté tous les verbes et les noms du corpus.

2.3 Articles des extractions des évènements :

2.3.1 Une approche TF-IDF et basée sur la co-occurrence pour l'extraction d'événements du corpus des nouvelles arabe :

Description

Une approche basée sur **le calcule de TF-IDF** pour extraire des mots-clés À partir les titres d'articles de presse Arabe. Ces mots-clés serviront à extraire les principaux événements pour chaque mois en utilisant une approche basée sur la cooccurrence de la partie du discours (POS).

L'approche utilisée

Ils ont utilisé un corpus de textes d'actualité pour extraction des évènements, cette approche considérer que chaque mot spécifique pourrait être liée à un évènement spécifique en général, les principaux textes d'actualité contiennent des phrases hétérogènes qui peuvent être liées ou non à l'événement. Pendant ce temps, les titres sont représentés par une seule phrase qui décrit le nouvel article. Ce système extraire les mots-clés présentés par les mots les plus répétés des titres. Le système extraire les cooccurrences des entités nommées avec des mots-clés filtrés comme le montre la Figure 2.3.

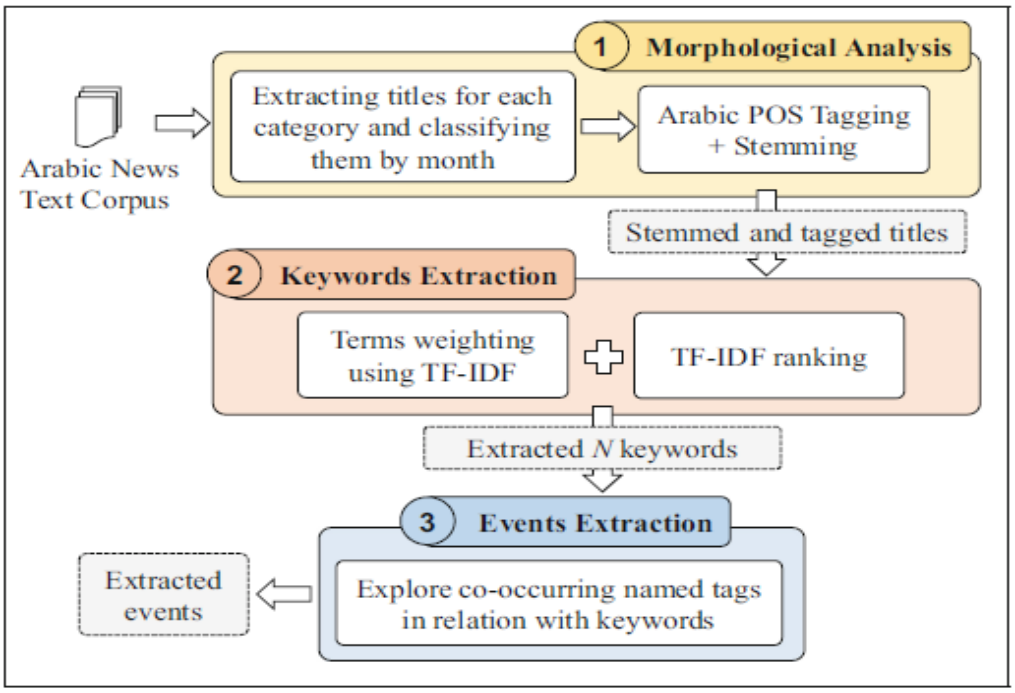

Figure 2.3: La technique utilise les trois étapes suivantes [1]

a. Analyse morphologie : Après avoir extrait et classé les titres par mois le système utilise le tagger OpenNLP[ii] pour annoter les nouveaux titres avec l'étiquette grammaticale appropriée pour chaque mot.

ترامب يقيد الهجرة من هذه الدول العربية

(Trump restricts immigration from these Arab countries)

ترامب_B-PER يقيد_V الهجرة_D+SFN من_P هذه_DM الدول_D+PIN العربية_D+AJ

Figure 2.4: Exemple d'étiquetage grammatical [1].

b. Extraction les mots-clés : Après que le système affecte le POS, il calcule le score pour les noms extraits. La mesure TF-IDF est souvent utilisée comme facteur de pondération et pour le filtrage des mots vides dans les champs d'extraction d'informations et de texte. Afin le système d'identifier les mots clés les plus importants pour chaque mois.

c. Extraction d'événements basée sur la cooccurrence POS : Pour extraire les événements des titres le système il cherche les mots-clés dans un titre, Ensuite, ils recherché des noms co-occurrents, basés sur Balises POS à droite et à gauche du mot clé correspondant [1]. Comme le montre la Figure 2.5.

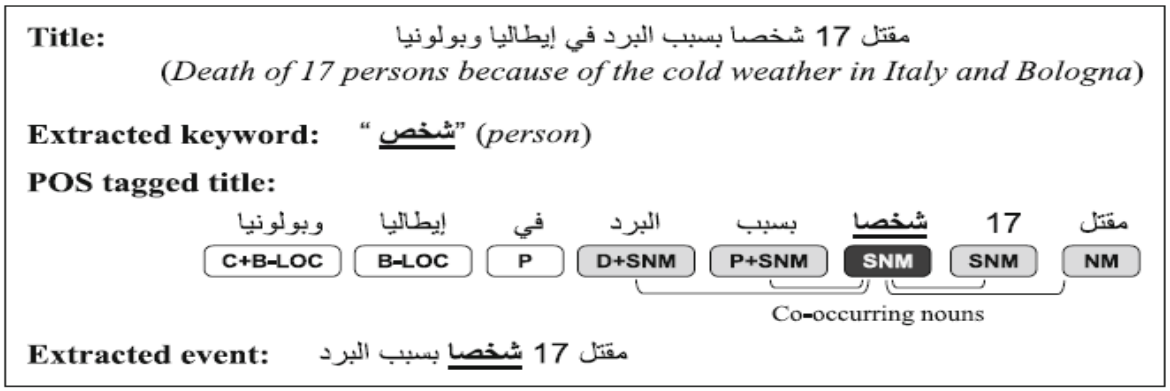

Figure 2.5: Un exemple expliquant le processus d'extraction d'événement [1].

Discussion

Les résultats dépendent également des catégories et fonctionnent bien pour des domaines spécifiques tels que l'économie. Car La catégorie sport a la valeur de précision la plus basse. Cette faible valeur pourrait s'expliquer par la richesse sémantique de cette catégorie. Les valeurs de précision sont calculées.

2.3.2 Approche basée sur la connaissance pour l'extraction d'événements à partir de tweets arabes

Description

Une **approche non supervisée basée sur des règles** pour l'extraction d'événements à été développé, et un système d'homonymie d'entité nommée à mapper chaque entité mention à leurs entités correspondantes qui sont représentés dans la base de connaissances.

Dans cette approche une base de connaissances ontologie a été conçue pour représenter les évènements extraits. Elle relie des entités de tweets avec des entités correspondantes sur les bases : Wikipédia, DBpedia, YAGO et Freebase). En plus, pour relier les évènements avec des dates en utilisant l'ontologie OWLtime.

L'approche utilisée

Pour chaque tweet, l'approche proposée extrait des entités nommées associées à leurs phrases d'événements et à leurs arguments d'événements spécifiques. L'expression des évènements peut s'agir des arguments suivants :

- *Agent d'évènements* : représente les acteurs de l'évènement.
- *Agent de localisation* : se réfère à la ville, pays ou au continent.
- *Agent cible* : La ou l'évènement est ciblé (son nom à titre d'exemple).
- *Déclencheur d'évènement* : Expression linguistique qui se réfère aux expressions des évènements.
- *Produit de l'évènement* : Parfois l'évènement annonce un produit.
- *Heure de l'évènement* : l'expression qui indique la date/temps de l'évènement.

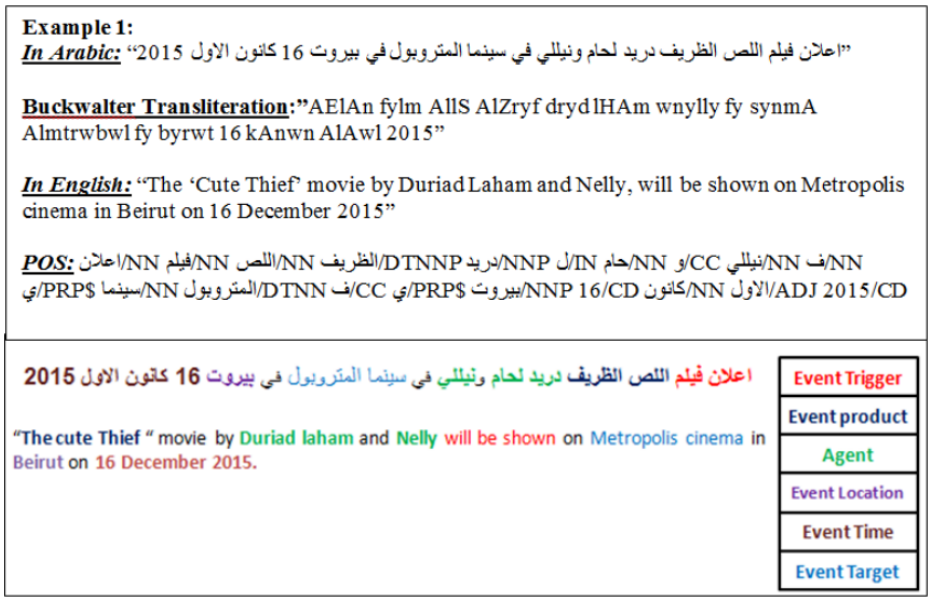

Figure 2.6: Exemple d'un tweet en arabe et sa POS [2].

a. Le prétraitement utilise un paquet AraNLP qui contient des services de tokezination et de normalisation de texte sous JAVA, marquage POS, et stemming. Dans cette phase les tweets sont nettoyés (élimination des mots non arabes, les hyper links, les hash-tags ainsi que les symboles). Le texte est donc normalisé en remplaçant chaque lettre par sa racine abstraite.

b. Extraction évènement et de désambigüité : Ce processus contient 3 sous processus :

i. Extraction des évènements : Pour ce faire, une approche basée sur des règles a été utilisée afin d'extraire les déclencheurs, le temps ainsi que le type d'évènement (instant ou intervalle).

ii. Entité nommée et désambigüité : il faut détecter le lien de l'agent, le produit et le lieu d'évènement tout sa on façon une relation entre ces EN du texte et qui corresponde dans les bases de connaissances.

iii. Relation temporelle : La liaison d'entités est utilisée pour résoudre les expressions temporelles extraites des tweets. Pour chaque événement extrait, les arguments d'événement sont liés à leurs correspondants dans la base de connaissances. Un nouvel événement est rempli dans la base de connaissances si et seulement si ses arguments n'y sont pas déjà représentés.[2]

Discussion

Les résultats de l'approche basée sur la connaissance pour extraire des événements des tweets arabes. Montrent que le système a une précision de, 75,9% pour T1 : extraction du déclencheur d'événement, 87,5% pour T2 : extraction du temps d'événement, et 97,7% pour T3 : identification du type d'événement.

2.3.3 Une approche hybride pour l'extraction d'événements

Desription

En ce qui concerne les approches hybrides, on cite le travail de Anup Kumar Kolya et al. Ils ont présenté ces travaux sur l'extraction d'événements. Au départ, ils ont développé un système supervisé basé sur CRF pour les événements d'extraction. Ces systèmes souffrent principalement d'identifier les verbes qui dénotent les expressions des événements. Par la suite, ils ont présenté plusieurs propositions afin d'améliorer les performances du système, puis ont proposé un nombre de techniques basées sur SRL, WordNet et les règles sémantiques.

L'approche utilisée

Ils ont utilisé trois approches pour extraire les événements :

a. Approche basée sur CRF (Conditional Random Field). 6 caractéristiques ont été utilisées.

b. Utilisation des règles sémantiques : Le but de cette étape est d'identifier les différentes caractéristiques des phrases d'un document. Ces fonctionnalités n'aident qu'à extraire les événements du texte.
c. Utilisation de WordNet : Les fonctionnalités de WordNet ont été largement utilisées pour extraire différentes catégories lexicales, telles que partie de discours (POS), stem, hypernym, meronym. WordNet est principalement utilisé pour identifier les noms d'événements non-verbaux.
d. Utilisation de règles pour l'extraction d'événements : Dans cette étape, ils sont concentrés principalement pour identifier les classes lexicales spécifiques.

Initialement, ils ont exécuté CRF basée sur Stanford, l'entité nommée (EN) sur l'ensemble de données de test TempEval-2. La sortie du système est étiquetée avec : Personne, Lieu, Organisation et Autres classes. [3]

Discussion

Dans cet article, ils ont présenté des travaux sur l'extraction d'événements dans le cadre de l'exercice d'évaluation TempEval-2010. Initialement, ils ont développé un système supervisé basé sur CRF pour l'extraction d'événements. Ces systèmes basés sur le CRF souffrent principalement de l'identification des noms déverbaux qui désignent les expressions de l'événement. Par la suite, ils ont présenté plusieurs propositions visant à améliorer les performances du système. Ils ont proposé un certain nombre de techniques basées sur les règles, WordNet et des règles manuelles. Les résultats de l'évaluation donnent les valeurs de précision, de rappel et de mesure F de 93,00%, 96,00% et 94,47%, respectivement.

2.4 Comparaison :

Dans le tableau suivant, nous comparerons les différents travaux suscités :

Tableau 3: Tableau comparatif des travaux d'extraction

Référence	Approche	Corpus	Type Corpus De corpus	Méthode	Rappel	Précision	F-mesure
Hassina Aliane[5]	Approche non supervisée	Corpus des articles WEB	Annoter	Utilisant des déclencheurs et des réglés			84%
Khaled Shaalan et Hafsa Raza[6]	Approche à base de régales	ACE[1] et ATB[2]	Annote	Utilisant des grammaires	96 ,8%	77 ,4%	85 ,9%
Iman Saleh,Lamia Tounsi et Josef van Genabith[4]	Apprentissage Automatique	ACE et ATB	Annote	Support vecteur machine			73%
Anup Kumar Kolya, Asif Ekbal et Sivaji Bandyopadhyay[3]	Hybride	TempEval-2010	Annote	Champ aléatoire conditionnel Et Base de connaissance	96%	93%	94,47%
Amina Chouigui, Oussama Ben Khiroun et Bilel Elayeb[1]	Approche statistique	Corpus v1.1[3]	Non annote	TF-IDF ET cooccurence		75,3%	
Mohammad AL-Smadi et Omar Qawasmeh[2]	Approche non supervisée	Twitter Streaming	Non annote	A base de règle		75,9%	

2.5 Conclusion

Après consultation des différents travaux d'extraction d'évènements et d'expression temporelles, une étude comparative a été faite afin de choisir une approche adéquate pour notre travail. On a conclu que la meilleure approche est bien l'apprentissage automatique. Malheureusement, par manque de corpus, celle-ci est mise à l'écart. L'approche sur les bases de connaissances (dictionnaire) ne peut être efficace à cause de richesse de la langue arabe. On choisira donc l'approche à base de règle avec un choix d'indice linguistique robuste (étiquetages grammatical). Cette dernière sera bien détaillée dans le chapitre suivant.

[1] http://www.ircs.upenn.edu/arabic
[2] http://projects.ldc.upenn. edu/ace
[3] https://antcorpus.github.io

PARTIE 2 : CONCEPTION, IMPLEMENTATION ET EVALUATION

Chapitre 3. Conception

3.1 Introduction

Par définition l'extraction des évènements et des expressions temporelles est la tâche d'identifier un mot ou bien une suite de mots d'un article permettant de signaler les évènements ainsi que la date de ce dernier. Un system d'extraction est jugé efficace s'il permet d'extraire l'évènement principal et son successeur.

Notre travail est de concevoir un système d'identification applicable sur des textes arabes. Ces textes sont des articles écrits en arabe standard, moderne et de différentes catégories.

Pour ce faire, notre système suit tout un processus où il débute par une étape de segmentation et filtrage en passant par les traitements linguistiques (normalisation, étiquetage grammatical, les entités nommées) pour enfin appliquer les grammaires (automates) et les règles linguistiques.

3.2 Caractéristiques du corpus

Le corpus utilisé est l'ANT Corpus v1.1. Ce dernier contient des articles de presse de janvier à juin 2017 classés par domaine. Les textes ANT Corpus sont collectés à partir du site tunisien de la radio web Jawhara FM.

3.3 L'Approche propose

En général, pour la langue arabe, les approches servant à extraire les évènements et les expressions temporelles dans la littérature ne sont implémentées que dans les tweets ou pour des petites phrases où l'on ne trouve qu'un seul évènement. Par conséquent, dans le cas des articles arabes où l'on trouve plusieurs évènements et plusieurs types d'expressions temporelles, les approches sont rares voire inexistantes.

Dans notre cas, on s'est penchés pour une méthode « à base de règles ». Ce choix s'est fait principalement par rapport à son efficacité à extraire les évènements et les expressions

temporelles dans les articles arabes où le verbe est généralement considéré comme déclencheur d'évènements ; En plus de la disponibilité d'outils d'indices linguistiques.

Pour l'extraction des évènements, l'idée est inspirée du travail de **Mohamed [5]**. Mais son approche n'est pas efficace en présence des adverbes de négations avant le verbe. Ce qui nous a poussés à perfectionner ce dernier.

Le travail de **SHAALAN, Khaled et RAZA, Hafsa [9]** est intéressant en ce qui concerne les expressions temporelles. Néanmoins, ce dernier touche des formes et des types bien précis. On veut donc l'élargir de telle sorte à toucher plus de formes et de types.

3.4 Architecture de système

L'architecture globale de notre système d'indentification d'évènements, de sa date et de son lieu, est représentée dans la Figure 3.1.

Le processus de l'application de ce système suit une série de traitements, en commençant par un filtrage et d'une segmentation suivie d'une normalisation puis un étiquetage grammatical et extraction d'entités nommées, pour enfin appliquer les règles et les grammaires (automates).

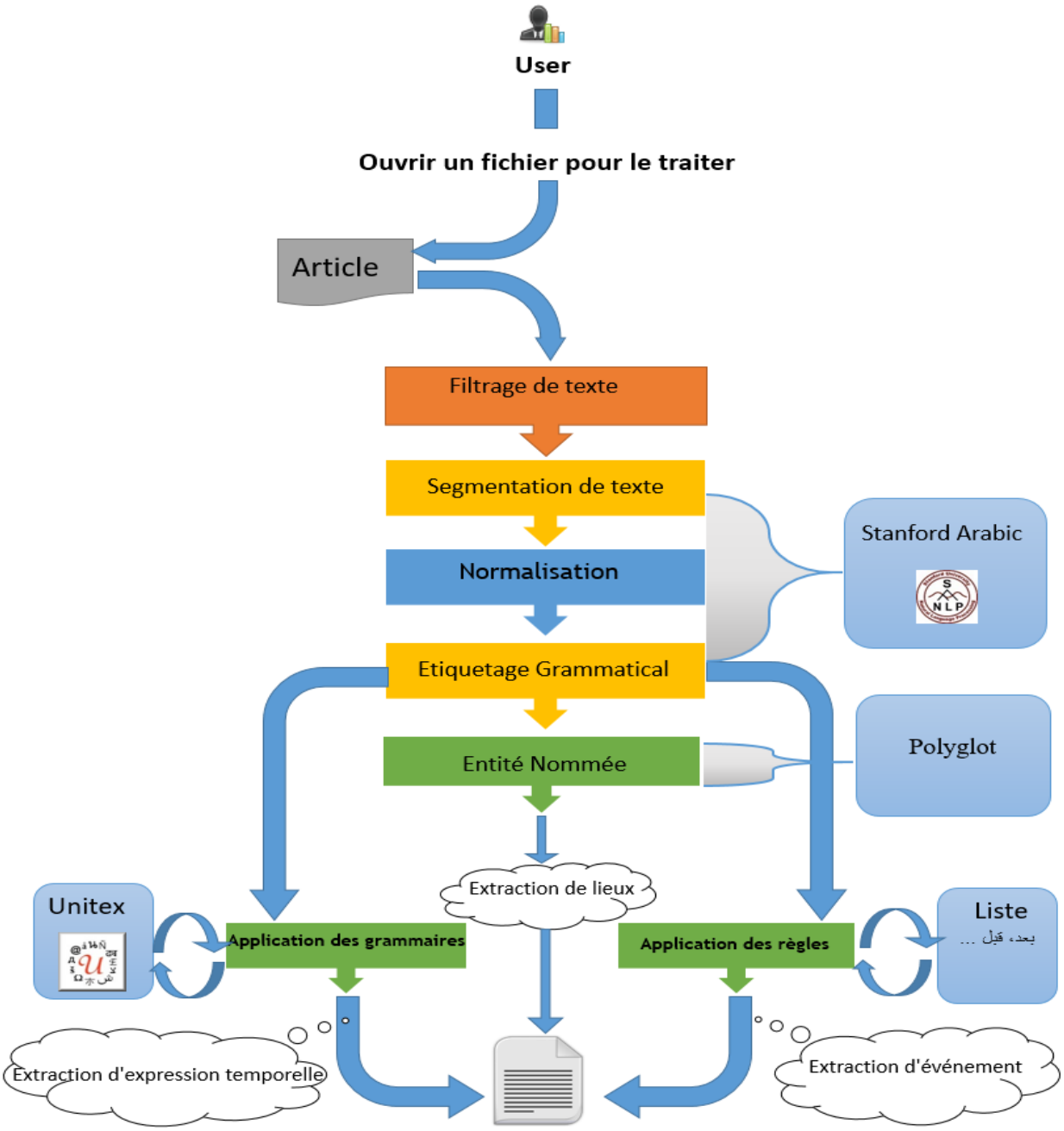

Figure 3.1: Architecture globale.

3.4.1 Prétraitements

Filtrage manuel des articles

Le filtrage manuel des articles consiste en la suppression des figures et des schémas inutiles, car on ne traite que des fichiers textuels.

Encodage uniques des textes

Tous les inputs de notre système doivent être représentés dans un fichier texte .TXT avec encodage UTF-8. Pour qu'il n'est pas de déformations au niveau des caractères lors de sa lecture.

Cette étape permet donc de convertir un article de type (.doc) fichier word a un fichier texte (.txt).

3.4.2 La segmentation :

Cette étape permet de découper les termes du texte en segments, de sorte qu'ils puissent être introduits dans un capteur morphologique ou dans un étiqueteur de position pour un traitement ultérieur.

Dans notre approche nous avons utilisé le segmenteur de la bibliothèque de **StanforSegmenter**[4]

[4] https://stanfordnlp.github.io/CoreNLP/ (consulté le 25 juin 2019)

qui sert à segmenter les mots composés en tokens.

Exemple :

تقدمت كتلة حرة بمجلس النواب الشعب تطلب فيها رفع قضية

تقدمت كتلة حرة ب مجلس النواب الشعب تطلب في ها رفع قضية

3.4.3 Analyse et traitement linguistique

La normalisation

Plusieurs genres de normalisation sur le texte sont appliqués afin de mieux manipuler les variations du texte qui peuvent être représentées en arabe. Par exemple, dans l'arabe écrit, les voyelles sont souvent omises dans les textes, néanmoins, on peut parfois trouver quelques voyelles présentées avec des mots. Alors, l'élimination de ces voyelles est nécessaire pour mettre fin à la normalisation. Certaines lettres subissent une simple modification dans l'écriture qui n'influe pas considérablement sur le sens du mot. Mais l'encodage de ces lettres change d'un mot à un autre. Une autre raison pour ce prétraitement est que l'on a tendance fréquemment à mal écrire ces différentes formes de hamza. Ce genre d'erreurs est très répandu dans les textes arabes. Par exemple, le mot « أكل « est généralement écrit « <اكل >.

La normalisation concerne les étapes suivantes :

¬ Retirer les signes diacritiques (principalement voyelles faibles).

¬ Remplacer le إ ou le أ initial par l'alif nu ا.

¬ Remplacer le آ par le ا.

Ces opérations sont faites par la fameuse bibliothèque **Stanford**

Etiquetage grammatical par StanfordTagger

Un tagueur de partie du discours (POS Tagger) est un traitement qui lit un texte et assigne une partie du discours à chaque mot : nom, un verbe, un adjectif, etc., bien qu'en règle générale les applications utilisent des balises POS. Et on a modifié les résultats obtenu par StanfordTagger afin qu'il soit plus simple comme il apparaît dans l'exemple suivant :

Exemple :

Résultat du StanfordTagger :

['تقدمت', 'VBD'], ['كتلة', 'NN'], ['الحرة', 'DTJJ'], ['ب', 'IN'], ['مجلس', 'NN']

Résultat de notre traitement :

['تقدمت', 'V'], ['كتلة', 'NN'], ['الحرة', 'DTJJ'], ['ب', 'IN'], ['مجلس', 'NN']

3.4.4 Application des règles et des grammaires :

Application des grammaires :

Pour extraire les expressions temporelles, on a implémenté grâce au logiciel Unitex des graphes (graphes des grammaires) qui traitent les différentes variétés de la langue arabe. Tout un programme a été implémenté (avec une autre interface présentée dans le chapitre qui suit) afin d'extraire toutes les expressions temporelles d'un texte donné. Etant donné que seule la première expression qui nous intéresse, ce dernier a été modifié de telle sorte à extraire que la première expression.

Pour extraire la première expression temporelle, on a utilisé le logiciel Unitex qui facilite l'implémentation des grammaires (ou bien des graphes) qui traite les différentes variétés de la langue arabe.

En langue arabe, les expressions temporelles ont plusieurs types et formes (voir les exemples ci-dessous). Dans notre cas on prend en considération les types suivants :

- Les quatre types pour définir les mois (شهور , شهور عربية , الشهور السريانية,الشهور الهجرية رومانية)
- Les types de numéro utilisés (1,١)
- Types de description de l'année numéro ou bien un mot relatif tel que سابق (précédent).

On implémente alors des grammaires (des graphes) pour les quatre types de mois défini au-dessus pour reconnaitre les formes suivantes (voir figure 3.2):

صفر/فيفري Ou مارس\ أذار ou bien simplement مايو

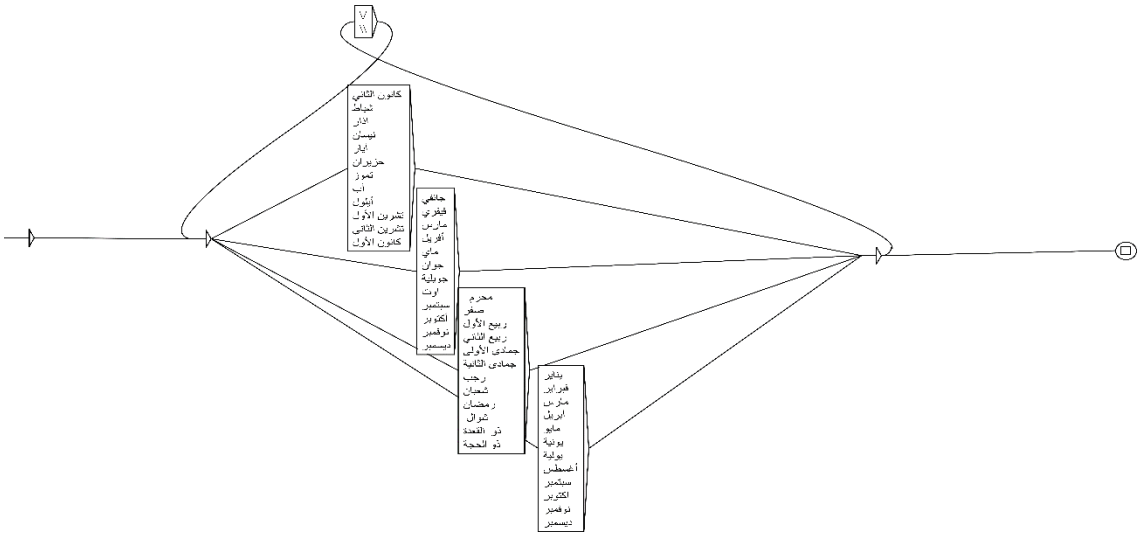

Figure 3.2: Graphe des différentes formes et types de mois

PS : les mois dans les graphes suivants seront mentionnés par le mot **MOIS** et dans les exemples par le mot **شهر**.

Ensuite on implémente quatre graphes qui utilisent le graphe principal des mois définis au-dessus pour les différentes formes possibles des expressions temporelles en arabe.

1er graphe :

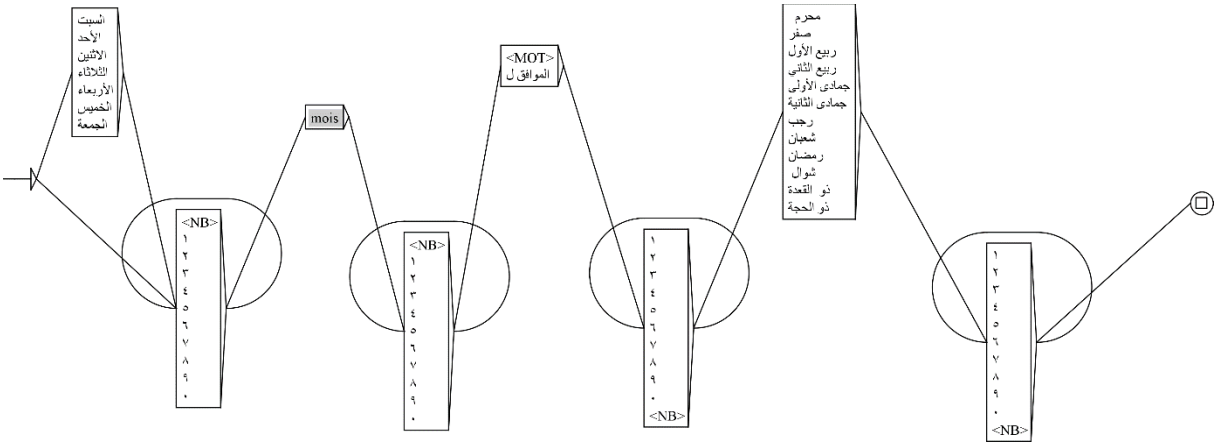

Figure 3.3: Graphe de forme type 1

Exemple : les formes reconnues par ce graphe sont :

السبت 20 شهر 2018 الموافق ل 14 صفر 1550

الاحد 10 شهر 2015 الموافق ل ١٢ رمضان ١٣٤٥

20 شهر 2018 الموافق ل 14 صفر 1550

10 شهر 2015 الموافق ل ١٢ رمضان ١٣٤٥

2eme graphe :

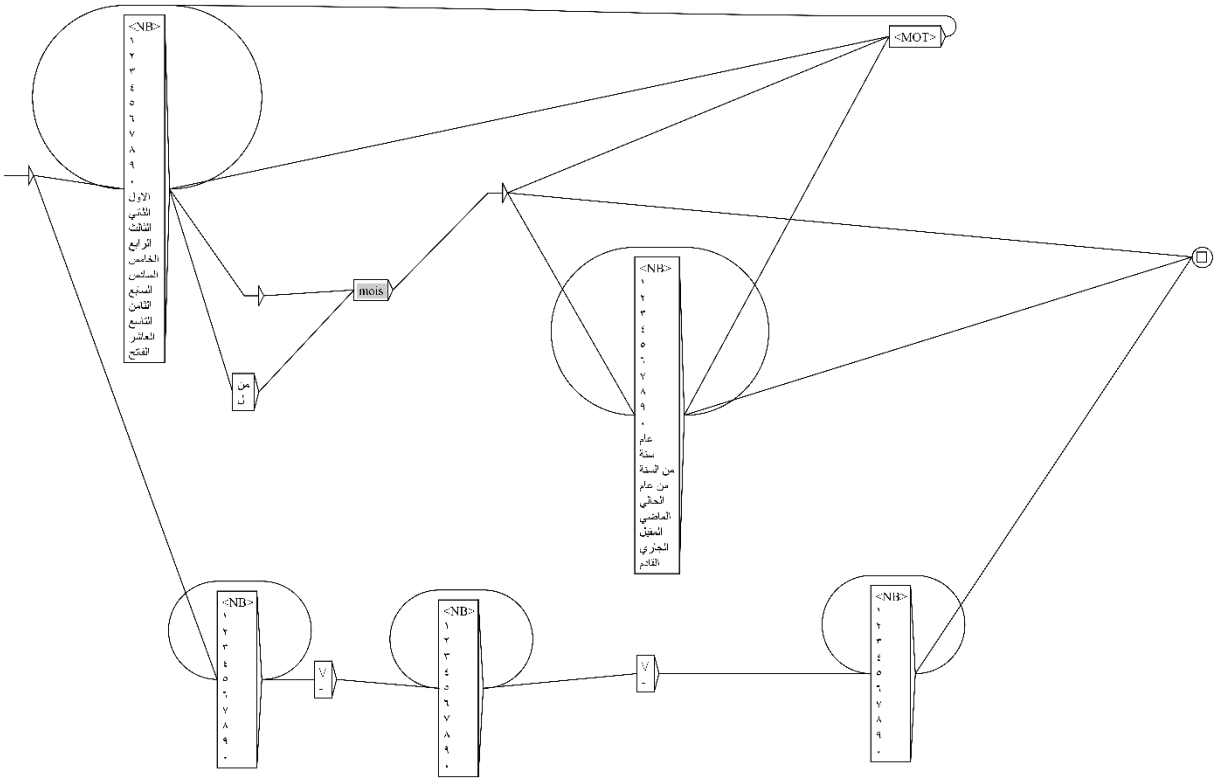

Figure 3.4: Graphe de forme type 2

Exemple : les formes reconnues par ce graphe sont :

20 الى 30 شهر(2018/القادم/عام 2018/من سنة 2018...)حتى 20 شهر (2018/القادم/عام 2018/من سنة 2018...)

20 شهر(2018/القادم/عام 2018/من سنة 2018...) الى 15 شهر (2018/القادم/عام 2018/من سنة 2018...)

20 شهر الى 25 شهر(2018/الماضي/عام 2018/من سنة 2018...)

20 الى 15 شهر (2018/الحالي/عام 2018/من سنة 2018...)

1994/06/30

1994-06-30

3eme graphe :

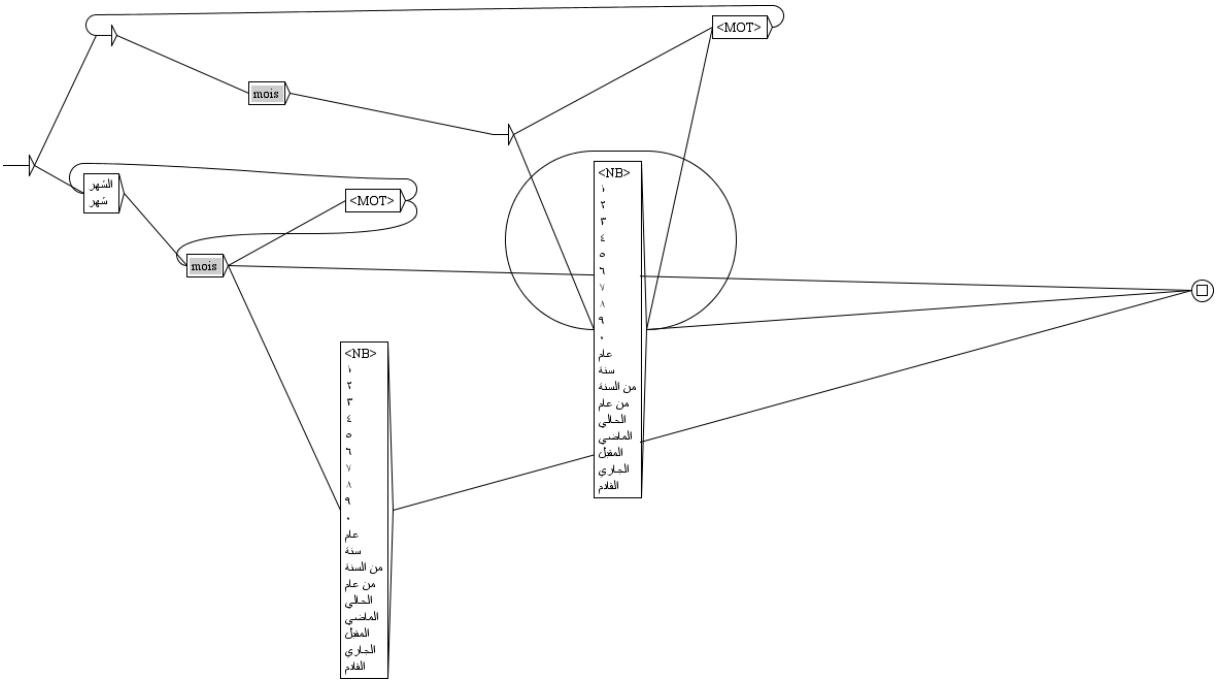

Figure 3.5: Graphe de forme type 3

Exemple : les formes reconnues par ce graphe sont :

شهر(2018/الحالي/عام 2018/من سنة 2018...) الى شهر (2018/الحالي/عام 2018/من سنة 2018...)

شهر الى شهر (2018/الحالي/عام 2018/من سنة 2018...)

شهر الى شهر

4 eme graphe :

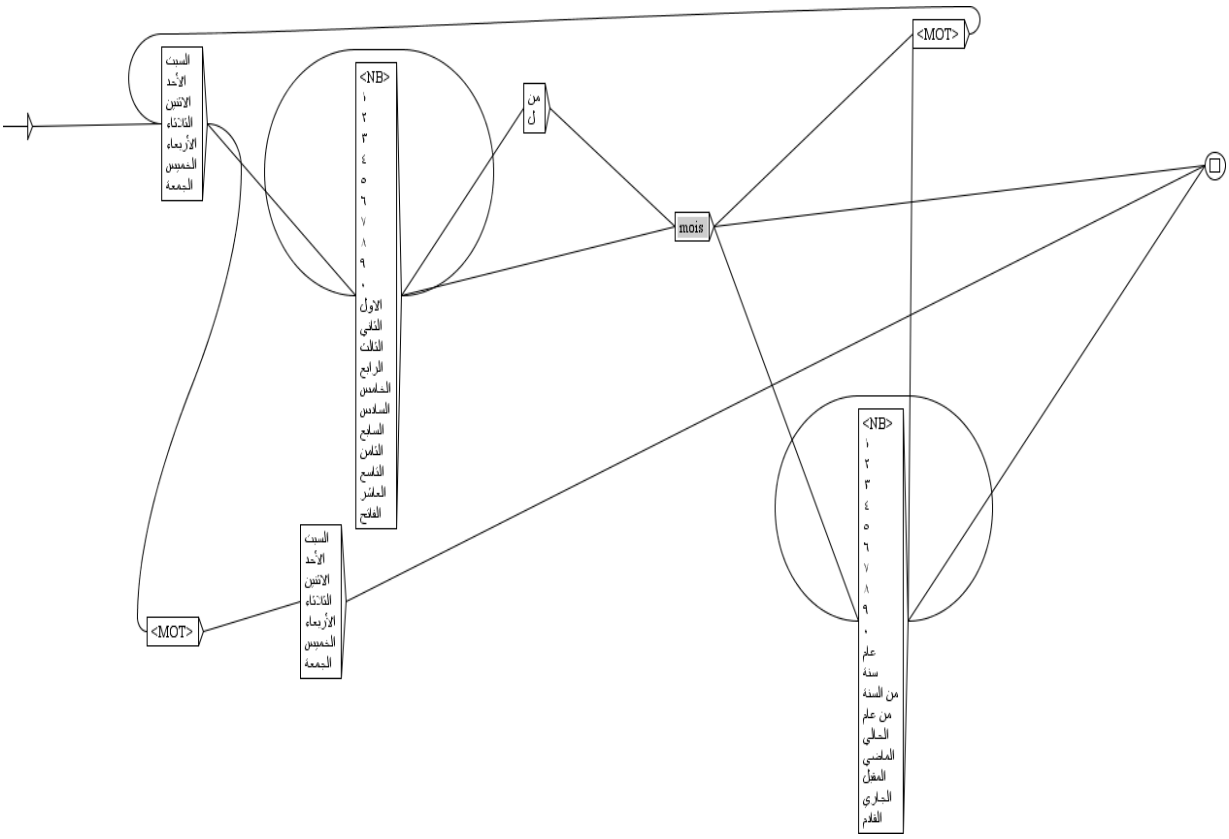

Figure 3.6: Graphe de forme type 4

Exemple : les formes reconnues par ce graphe sont :

السبت (25/الأول/٢٥/الفاتح) شهر(1950/الحالي/عام 1989/من سنة 1999) الى الأحد (25/الأول/٢٥/الفاتح) شهر (2000/الحالي/عام 2000/من سنة 2000)

السبت 25 شهر الى الأحد 10 شهر 2005

السبت 25 شهر الى الأحد 10 شهر

السبت الى الخميس شهر 2010

السبت الى الخميس شهر

السبت الى الخميس

Identification des lieux

Polyglot [5] *utilisé pour identifier les lieux dans un article. En prenant le premier lieu apparent.*

Exemple :

انفجار قنبلة في فرنسا يوم السبت

Application des règles

Afin d'extraire les évènements, on applique des règles sur l'étiquetage des termes de l'article pour les deux types de phrase :

a) Phrase verbale

Ce type de phrase commence toujours par un verbe (V). Ce dernier est considéré comme le premier évènement jusqu'au prochain nom (NN) ou jusqu'à ce que la suite de mots soit inférieure à cinq. Ensuite on vérifie s'il y a un autre verbe (V) pour réappliquer la même règle, mais en ajoutant la négation si elle existe (RP) (...لا,لن).

Exemple :

أعلنت وزارة الدفاع الجزائرية أنه لا يمكن اجراء الانتخابات.

أعلنت وزارة الدفاع الجزائرية أنه لا يمكن اجراء الانتخابات

b) Phrase nominale

Ce type commence par différents noms. On distingue trois cas :

- Phrases nominales avec au début un mot de liste noire et contenant verbe(s):

Cette phrase commence par un mot qui figure dans notre liste noire.

[5] https://polyglot.readthedocs.io/en/latest/ (consulté 28 juin 2019)

بعد أيام ... تجرى .... / قبل قليل أعلنت الرئاسة

Dans ce cas on ignore ces mots et on applique la règle de la phrase verbale.

- Phrases nominales qui ne contiennent ni de mots de liste noire au début ni verbe :

  Dans ce cas-là, on prend tous les mots (maximum 4 mots), car la phrase est petite.

Exemple :

عطلة عيد الأضحى يوم السبت

- Phrases nominales contenant verbe(s) :

On prend tous les mots qui précèdent le verbe, puis on vérifie s'il existe un autre verbe afin d'appliquer la deuxième partie de la phrase verbale.

Exemple :

وزارة الدفاع الجزائرية أعلنت أنه لا يمكن اجراء الانتخابات

3.5 Conclusion

Dans ce chapitre, on a décrit notre application, dont l'objectif était de concevoir un système capable d'identifier automatiquement un évènement avec sa date et son lieu à partir d'articles arabes. La conception de système est divisée en 5 étapes. Cette dernière sera mise en fonction dans le chapitre qui suit.

Chapitre 4. Implémentation

4.1 Introduction

Nous traiterons dans ce chapitre la mise en œuvre de notre système.
D'abord, on présentera l'environnement de développement et décrira le langage de programmation utilisé. Ensuite, nous expliquerons les progrès de notre application en mentionnant les bibliothèques utilisées, pour enfin décrire les étapes de la progression.

4.2 Environnement de développement

Dans Cette section on présente le langage de programmation Python utilisé, l'environnement de développement PyCharm et le logiciel Unitex.

4.2.1 Python

Python est un langage de programmation interprété orienté objet. C'est un langage de programmation puissant et facile à apprendre. Il possède des structures de données de haut niveau efficaces et une approche simple mais très efficace de la programmation orientée objet.
Il a été créé par Guido van Rossum, un ancien résident des Pays-Bas, dont le groupe de comédie préféré à l'époque était Flying Circus de Monty Python. Il a été conçu à la fin des années 1980 comme un successeur de la langue ABC. Python 2.0, publié en 2000, a introduit des fonctionnalités telles que la compréhension de liste et un système de récupération de place capable de collecter des cycles de référence. Python 3.0, publié en 2008, était une révision majeure du langage qui n'était pas totalement compatible avec les versions antérieures, et une grande partie du code Python 2 ne s'exécute pas sans modification sur Python 3.
La syntaxe élégante et le typage dynamique de Python, ainsi que sa nature interprétée, en font un langage idéal pour les scripts et le développement rapide d'applications dans de nombreux domaines de la plupart des plates-formes.

C'est l'un des langages de programmation les plus populaires pour les scientifiques et les éducateurs. Il est largement utilisé en informatique scientifique industrielle.
Il est dit qu'il est relativement facile à apprendre et portable, ce qui signifie que ses instructions peuvent être interprétées dans plusieurs systèmes d'exploitation, notamment les systèmes UNIX, Mac OS, MS-DOS, OS / 2 et diverses versions de Microsoft Windows.
Python est couramment utilisé dans les projets d'intelligence artificielle avec l'aide de bibliothèques telles que Ten-sorFlow, Keras et Scikit-learn. Langage de script à l'architecture modulaire, à la syntaxe simple et aux outils de traitement de texte enrichi, Python est souvent utilisé pour le traitement du langage naturel[11].

4.2.2 PyCharm

PyCharm est un environnement de développement intégré (IDE) utilisé pour la programmation en Python. Il fournit une analyse de code, un débogueur graphique, un testeur d'unité intégré, une intégration aux systèmes de contrôle de version (VCS) et prend en charge le développement Web avec Django. Il est développé par la société tchèque JetBrains.
Il fonctionne sur plusieurs plates-formes Windows, Mac OS X et Linux. PyCharm a une édition professionnelle et une édition communautaire.
Il fournit la complétion intelligente de code, les inspections de code, la mise en évidence à la volée des erreurs et des solutions rapides, ainsi que la refactorisation automatique du code et des fonctionnalités de navigation avancées.
Il s'intègre à IPython Notebook, possède une console Python interactive et prend en charge Anaconda ainsi que plusieurs packages scientifiques, notamment matplotlib et NumPy.
Outre Python, PyCharm prend en charge les langages JavaScript, CoffeeScript, TypeScript, Cython, SQL, HTML / CSS, les langages de gabarit, AngularJS, Node.js, etc.
Il a de nombreuses fonctionnalités telles que :

- Assistance et analyse du codage, avec complétion du code, mise en évidence de la syntaxe et des erreurs, intégration de l'interface et solutions rapides.
- Navigation de projet et de code : vues de projet spécialisées, vues de structure de fichiers et sauts rapides entre fichiers, classes, méthodes et usages.

• Refactoring Python : inclut renommer, extraire une méthode, introduire une variable, introduire une constante, tirer vers le haut, pousser vers le bas, etc. Prise en charge des infrastructures Web : Django, web2py et Flask.

• Débogueur Python intégré.

• Test unitaire intégré, avec une couverture ligne par ligne.

• Développement Python de Google App Engine.

• Version Control Integration: interface utilisateur unifiée pour Mercurial, Git, Subversion, Perforce et CVS avec listes de modifications et fusion.[12]

4.2.3 Unitex

Unitex est un ensemble de logiciels permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire. Elles sont issues de travaux initiés sur le français par Maurice Gross au Laboratoire d'Automatique Documentaire et Linguistique (LADL). Ces travaux ont été étendus à d'autres langues au travers du réseau de laboratoires RELEX.

Unitex est un moteur permettant d'exploiter ces ressources linguistiques. Ses caractéristiques techniques sont la portabilité, la modularité, la possibilité de gérer des langues possédant des systèmes d'écritures particuliers comme certaines langues asiatiques et l'ouverture, grâce à une distribution en logiciel libre. Ses caractéristiques linguistiques sont celles qui ont motivé l'élaboration des ressources : la précision, l'exhaustivité et la prise en compte des phénomènes de figement, notamment en ce qui concerne le recensement des mots composés. [13]

4.3 Description du système

Le langage utilisé dans notre application est le Python 3.6. Unitex est utilisé aussi pour exécuter les commandes servant à utiliser les graphes implémentés dans l'application. Le tout dans un environnement PyCharm qui permet d'utiliser les différentes API (Application Programming Interfaces) à partir de plusieurs langages de programmation tels que Java.

Nous avons utilisé quelques fonctions fournies par différentes bibliothèques, on cite :

Bibliothèque Stanford[6] : Stanford CoreNLP fournit un ensemble d'outils technologiques en langage humain. Il peut donner les formes de base des mots, leurs parties du discours, qu'il s'agisse de noms de sociétés, de personnes, etc., normaliser les dates, les heures et les quantités numériques, marquer la structure des phrases en termes de syntagmes et de dépendances syntaxiques, indiquer quelles expressions nominales font référence aux mêmes entités, indiquent un sentiment, extraient des relations particulières ou à classes ouvertes entre des mentions d'entités, obtiennent les citations que les gens ont dites, etc .

Les outils offerts par Stanford : la segmentation (StanfordSegmenter) , categorie gramatical (StanfordPOStagger) .

Bibliothèque Polyglot[7] : Polyglot est un pipeline de langage naturel prenant en charge d'énormes applications multilingues. Il prend en charge la tokenisation pour 165 langues, la détection de la langue pour 196 langues, la reconnaissance d'entités nommées pour 40 langues, l'étiquetage vocal pour 16 langues, l'analyse des sentiments pour 136 langues, les enveloppements de mots pour 137 langues, l'analyse morphologique pour 135 langues, la translittération pour 69 langues [2]. Dans notre programme on a utilisé la reconnaissance d'entités nommées pour la langue arabe offert par Polyglot.

D'autres bibliothèques sont aussi utilisées comme **WebBrowser**[8] pour les pages HTML, **BeautifulSoup**[9] pour le filtrage du corpus.

[6] https://stanfordnlp.github.io/CoreNLP/
[7] https://polyglot.readthedocs.io/en/latest/
[8] https://docs.python.org/2/library/webbrowser.html
[9] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

4.4 Déroulement

On présente dans cette section les différentes étapes de déroulement du processus d'extraction de notre système, depuis la sélection de textes jusqu'à l'identification des déclencheurs, date et lieux d'un évènement. En commençant par l'interface globale dans la figure qui suit :

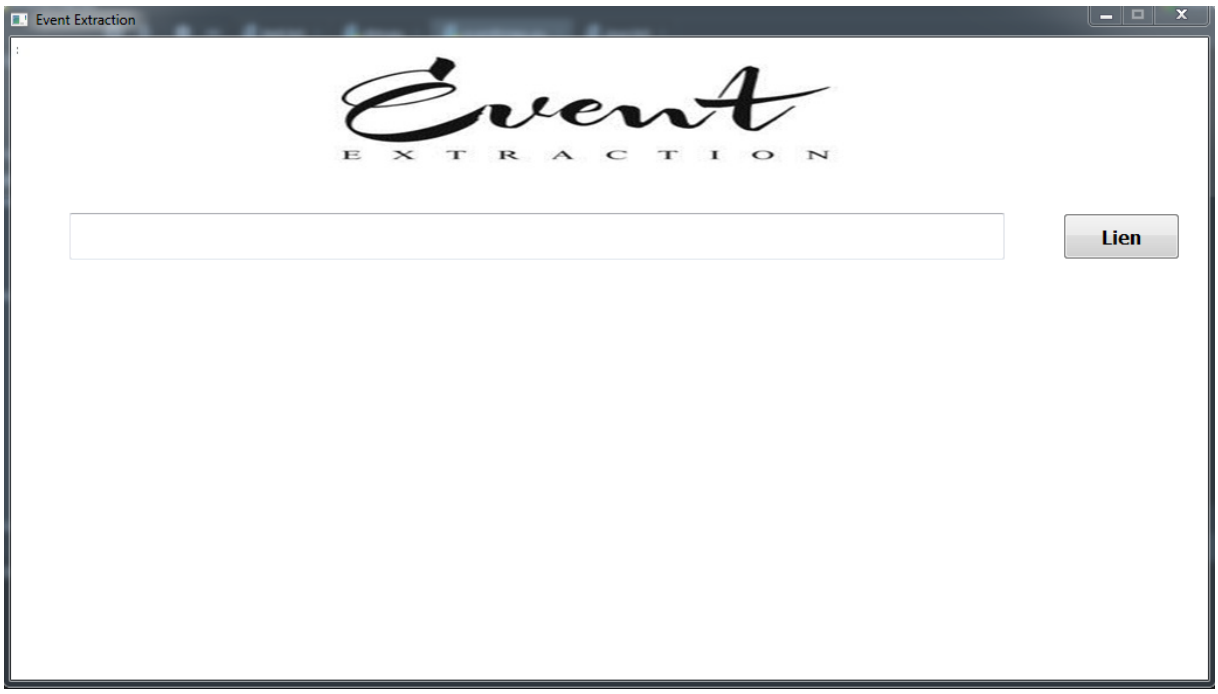

Figure 4.1: Interface globale

4.4.2 Sélection des textes

Les textes que nous avons utilisés sont de format « txt ». Le traitement automatique commence par la sélection d'un fichier à traiter, en activant le bouton **Lien**, et en choisissant le fichier a traité, dans la figure suivante :

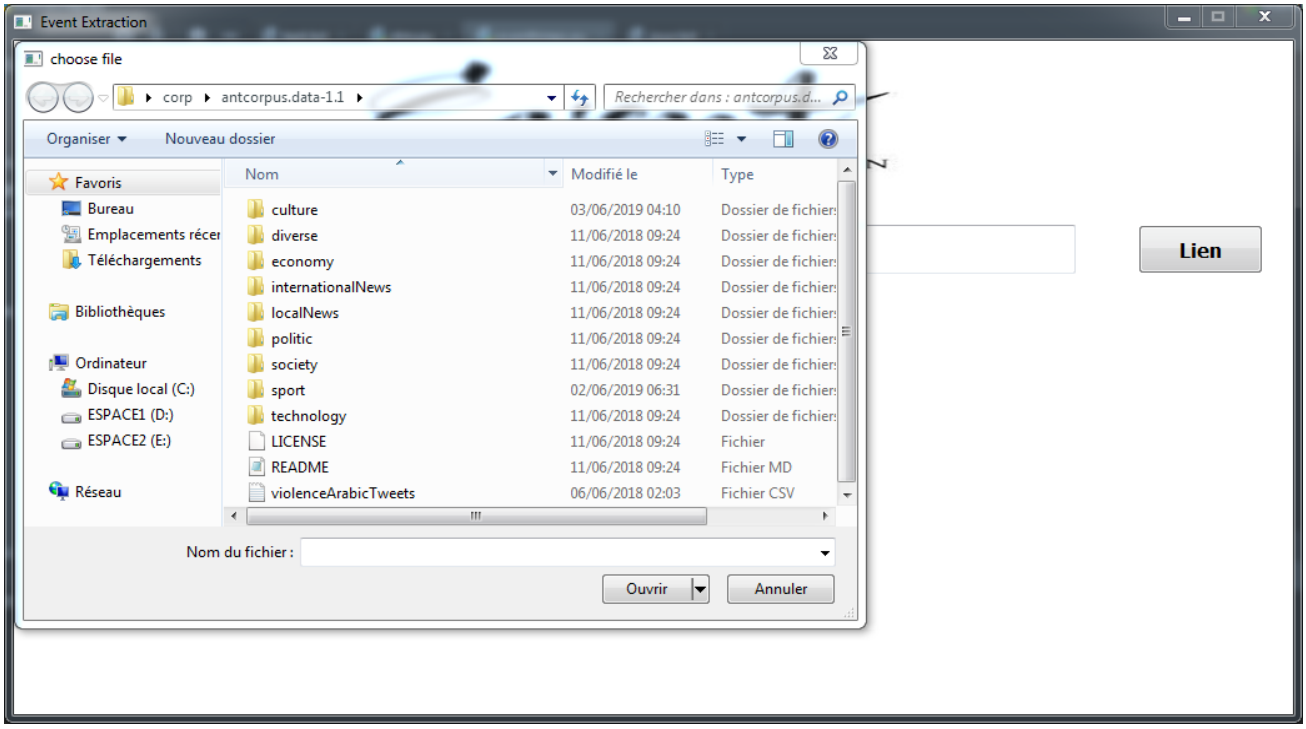

Figure 4.2: Sélection de fichiers

4.4.3 Choix du type d'expression temporelle

Après la sélection du fichier traité un choix de forme des expressions temporelles apparaît dans la comboboxe qui par défaut sélectionne « toutes les formes » voir la figure suivantes :

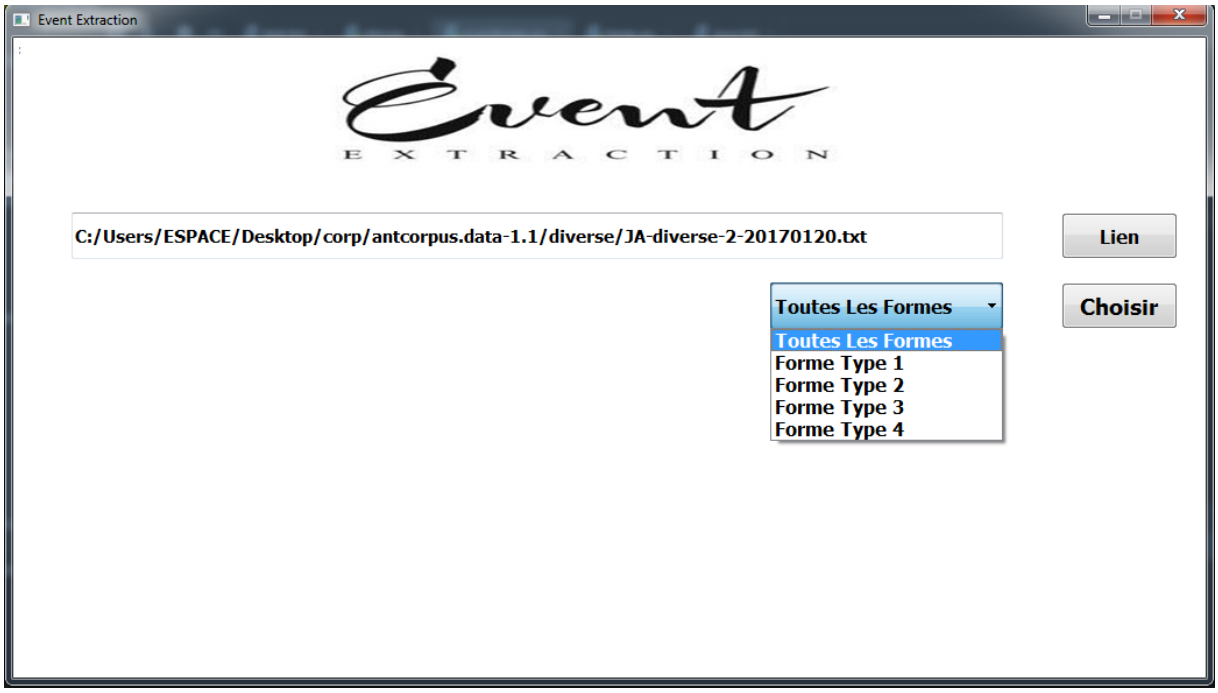

Figure 4.3: Sélection des formes

4.4.4 Exécution du processus

En appuyant sur le bouton **Choisir,** on a la main pour exécuter le programme **Exécuter.** On obtient alors les résultats **Premier Déclencheur, Deuxième Déclencheur, Date** et **Lieux** montrée en **figure 4.4.**

Dans cette étape on peut consulter le fichier traité grâce au bouton **Text** voir **Figure 4.5**.

L'option de refaire le travail est possible en appuyant sur le bouton A**ctualiser** comme le montre la **Figure 4.6**

Figure 4.4: Fenêtre de résultats d'exécutions

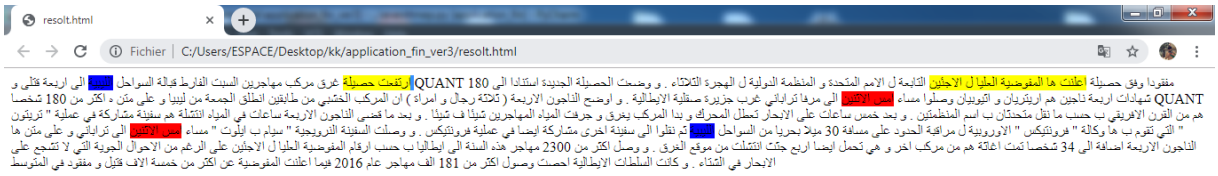

Figure 4.5: Consultation du texte traité

Figure 4.6: Fenêtre d'actualisation

4.4.5 Output de notre system

Dans le dossier d'application, on a un dossier qui contient les résultats d'exécution obtenus par l'Unitex (le texte traite dans un fichier html et un fichier des termes du texte).

4.4.6 Les expressions temporelles

Afin de bien exposer notre travail sur les expressions temporelles, on a implémenté une interface qui introduit un fichier .txt ainsi que extraire **Tout** les expressions temporelles possibles. Voir la figure suivante.

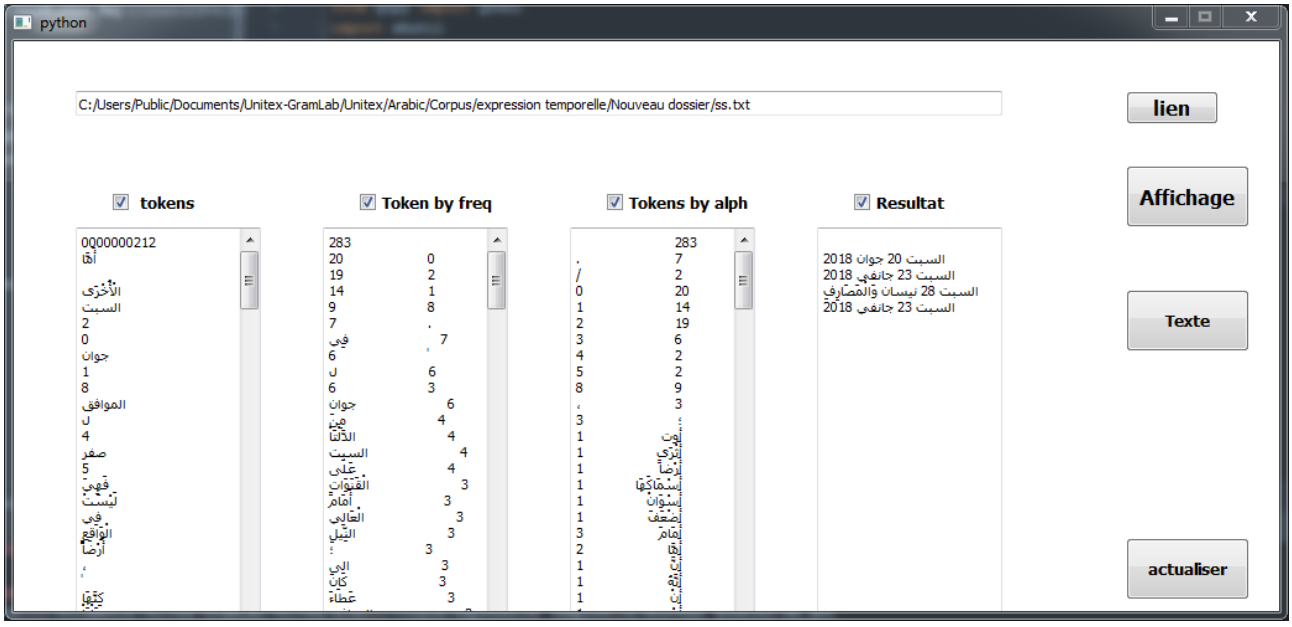

Figure 4.7: Fenêtre de résultats

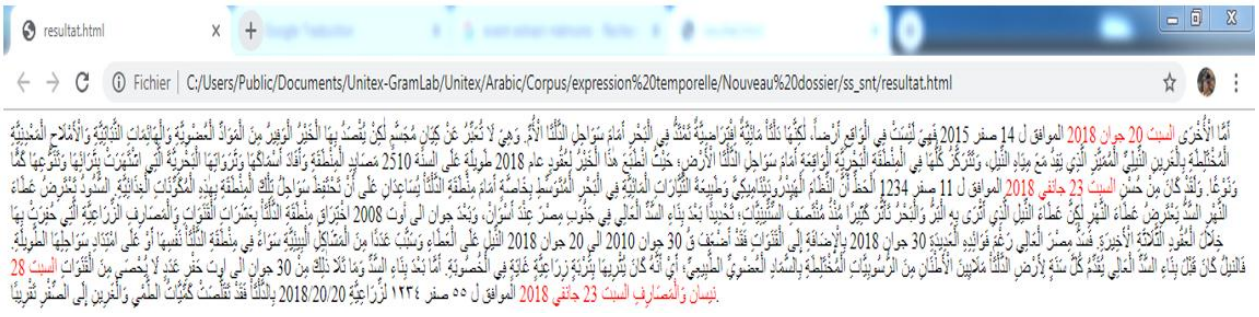

Figure 4.8: Consultation de texte traité

4.5 Evaluation du système

Pour évaluer la performance de notre système, nous avons fait une évaluation semi-automatique de ce dernier sur un ensemble de 30 articles de différentes catégories.

L'évaluation sert à comparer les évènements sélectionnés par le système et ceux définit préalablement par l'auteur. En utilisant les métriques Rappel et Précision afin de calculer F-mesure.

Les mesures Rappel, Précision et F-mesure sont calculées par les équations suivantes :

- $Rappel = Corrects / (Corrects + Oubliés)$
- $Précision = Corrects / (Corrects + Incorrects)$
- $F-mesure = 2.(Rappel.Précision) / (Rappel + Précision)$

Tel que :

***Correct*s :** Nombre d'évènement sélectionnés par le système et par l'auteur.

***Incorrects* :** Nombre d'évènement sélectionnés par le système et non pas par l'auteur.

***Oubliés* :** Nombre d'évènement sélectionnés par l'auteur et non pas par le système

Tableau

Cette petite comparaison montre l'efficacité de notre système d'identification. Par ailleurs, son évaluation est généralement subjective. Car le style de rédaction des textes diffère d'un auteur à un autre, et les formes ainsi que les types des expressions temporelles ne peuvent pas toutes être implémentées, à cause de la richesse de la langue. Cependant, le deuxième déclencheur reste un vrai problème.

Table 4.1: Résultats d'évaluation

Catégorie	Nombre de Document	Rappelle Evènement	précision Evènement	Rappelle Extraction d'expression temporelle	précision Extraction d'expression temporelle	Rappelle Extraction de lieux	précision Extraction de lieux
Culture	5	0,6	0,6	1	1	0,8	0,8
Economie	5	0,8	0,8	0,6	0,6	0,6	0,6
Politique	5	0,4	0,4	0,8	0,8	0,4	0,4
Sport	5	0,8	0,8	0,4	0,4	0,6	0,6
Technologie	5	0,2	0,2	0,8	0,8	0,6	0,6
Informations internationales	5	0,4	0,4	0,6	0,6	0,6	0,6

4.6 Conclusion

A travers ce chapitre, nous avons présenté l'environnement de développement de notre système ainsi que les différentes interfaces graphiques, qui à travers elles nous pouvons superviser le déroulement du système.

Ce système a pour rôle d'extraire les évènements à partir des articles de différentes catégories écrites en langue arabe. Les résultats de ce dernier sont liés étroitement au choix de l'approche implémentée.

L'utilisation des catégories grammaticales détectées par le StanfordPOS, et la modification des règles se sont avérées efficaces dans les textes que nous avons étudiés.

CONCLUSION

La problématique d'extraction des évènements et des expressions temporelles abordée dans ce mémoire s'est cristallisée autour de deux points : le premier concerne les critiques utilisées pour décider du contenu ou bien des termes essentiels à extraire. Le deuxième point se focalise sur le choix d'approche qui permet d'extraire ce contenu essentiel qui aide l'utilisateur dans ces traitements.

Pour comprendre et simplifier le sujet, nous avons présenté dans le premier chapitre le domaine d'extraction d'informations en général ainsi que l'extraction des évènements.

Dans le deuxième chapitre une étude est faite sur des articles afin de choisir l'approche utilisée. Malgré la rareté des travaux et la complexité de la langue arabe, on a pu améliorer une approche à base de règles faite pour des phrases courtes afin d'extraire les évènements, aussi des grammaires qui peuvent toucher plusieurs formes des expressions temporelles. Le tout détaillé dans le troisième chapitre.

Dans le chapitre quatre, des expérimentations et des évaluations du système sur un corpus qui regroupe un ensemble d'articles de différentes catégories écrites en arabe ont été faites afin de tester l'efficacité de notre système.

Dans le traitement automatique de la langue, la partie sémantique est essentielle. Malheureusement dans notre cas, le manque d'articles et des corpus (annotés) nous ont assiégés dans notre choix d'approche. Nous proposons donc, comme perspectives, de créer un corpus annoté afin d'élargir le choix.

Notre système s'avère peu efficace face à des formes d'expressions temporelles non traitées dans notre système, de même que quand il s'agit d'extraire les déclencheurs secondaires complexes. Ceci est dû à la négligence de la relation entre termes de la phrase.

En conclusion, ce travail montre que les approches appliquées sur les différentes langues sont difficiles à appliquer sur la langue arabe. Le traitement automatique de la langue arabe reste un domaine d'actualité et vierge qui reste un excellent sujet de recherche.

RÉFÉRENCES

[1] CHOUIGUI, Amina, KHIROUN, Oussama Ben, et ELAYEB, Bilel.A TF-IDF and co-occurrence based approach for events extraction from arabic news corpus.In : International Conference on Applications of Natural Language to Information Systems. Springer, Cham, 2018. p. 272-280

[2] MOHAMMAD, A. S. et QAWASMEH, Omar. Knowledge-based approach for event extraction from Arabic tweets.International Journal of Advanced Computer Science & Applications, 2016,
vol. 1, no 7, p. 483-490

[3] KUMAR KOLYA, Anup, EKBAL, Asif, et BANDYOPADHYAY,Sivaji.A hybrid approach for event extraction.Polibits, 2012,no 46, p. 55-59.

[4] SALEH, Iman, TOUNSI, Lamia, et VAN GENABITH, Josef. Zaman and raqm: extracting temporal and numerical expressions in arabic. In : Asia Information Retrieval

[5] ALIANE, Hassina, GUENDOUZI, Wassila, et MOKRANI, Amina.Annotating events, time and place expressions in arabic texts. In :Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. 2013. p. 25-31.

[6] SHAALAN, Khaled et RAZA, Hafsa. NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 2009,vol.60,no 8,p.1652-1663.

[7] RILOFF, Ellen et LORENZEN, Jeffrey. Extraction-based text categorization: Generating domain-specific role relationships automatically. In : *Natural language information retrieval*. Springer, Dordrecht, 1999. p. 167-196.

[8] Fabrice Even, Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale, Thèse de Doctorat, Université de Nantes, 05/10/2005

[9] Barigou Fatiha, Contribution à la catégorisation de textes et à l'extraction d'information, Thèse de Doctorat, Université d'Oran, 2012/2013.

[10] Y.C. Wu, T.K. Fan, Y.S. Lee, S.J Yen, "Extracting Named Entities Using Support Vector Machines", Spring-Verlag, Berlin Heidelberg, 2006

[11] https://docs.python.org/fr/3/tutorial/ (consulté le 15 juin 2019)

[12] https://www.jetbrains.com/pycharm/documentation/ (consulté le 16 juin 2019)

[13] PAUMIER, Sébastien, MARSCHNER, Sebastian Nagel, et STIEHLER, Johannes. UNITEX 3.2RC.