Extraction d'évènements au sein d'une plateforme de veille

Capucine Antoine

Extraction d'évènements au sein d'une plateforme de veille

ANTOINE

Capucine

Sous la direction de Thomas Lebarbé

UFR LLASIC

Département Sciences du Langage

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue – Orientation professionnelle

Année universitaire 2019-2020

Université Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : ANTOINE

PRENOM : Capucine

DATE : 03/09/2020

[redacted]

Mémoire de Master
Extraction d'évènements au sein d'une plateforme de veille

Remerciements

Je tiens tout d'abord à remercier mon encadrante *** *** pour son aide, sa rigueur, sa patience et ses conseils précieux. Elle est toujours restée à mon écoute et m'a permis d'acquérir de nouvelles connaissances.

Je tiens ensuite à remercier monsieur *** pour avoir accepté de m'accompagner dans l'aventure de ce stage et de m'avoir rassurée dans les moments stressants.

Je remercie également l'ensemble des professeurs du master IDL pour leur présence et leur gentillesse, et pour tous les projets intéressants qu'ils ont proposé à notre promotion de master, qui m'ont permis de découvrir le joyeux monde du TAL et de la programmation.

Je n'oublie pas *** *** , qui fut la meilleure binôme qui soit durant ces deux années de Master et qui m'a permis de progresser et de m'améliorer, notamment en programmation. Nous avons beaucoup travaillé, beaucoup stressé, souvent maudit nos scripts mais nous avons également énormément appris... Et beaucoup ri !

Résumé

L'extraction d'évènements est une application du Traitement Automatique des Langues (TAL), et plus précisément une application de fouille de textes, qui consiste à extraire de manière structurée des informations sur des évènements présents de manière non-structurée dans des textes. Lors de ce stage, nous avons travaillé à l'élaboration d'une maquette d'un outil d'extraction d'évènements qui repose sur une méthode de reconnaissance de motifs. Nous avons principalement travaillé avec le corpus ACE (Automatic Content Extraction) issu de la campagne d'évaluation du même nom. Ce stage s'inscrit dans le cadre de l'enrichissement continue des fonctionnalités d'AMI Enterprise Intelligence, la solution de veille stratégique développée et maintenue dans le centre R&D de la société Bertin IT à Montpellier.
Les principales contributions de ce mémoire sont :

- état de l'art sur l'extraction d'évènements ;
- étude de plusieurs outils d'extraction d'évènements disponibles en OpenSource ;
- développement d'une maquette d'un outil d'extraction d'évènements.

Mots-clés : Extraction d'évènements, Extraction d'information, Fouille de textes, Market Intelligence, Veille sur internet.

Abstract

Event extraction is an application of Natural Language Processing (NLP), more precisely of text mining, consisting in extracting structured information about events present in texts in an unstructured way. During the course of this internship, we worked on the development of the model version of an event extraction tool. This tool is based on a pattern-matching method. We mainly worked with the ACE (Automatic Content Extraction) corpus, from the evaluation conference of the same name. This internship and the developement of this model are part of enhancing AMI Enterprise Intelligence (AMI EI), the business intelligence solution developed and maintained by Bertin IT within its research center in Montpellier.
The main contributions of this project are :

- state of the art on event extraction ;
- review of several Open Source tools ;
- development of a model version of an event extraction tool.

Keywords : Event Extraction, Information Extraction, Text Mining, Market Intelligence, Web Intelligence.

Table des matières

Chapitre 1

Introduction

1 Présentation de l'entreprise

Société du groupe industriel CNIM (Constructions navales et industrielles de la Méditerranée), cotée à Euronext Paris, Bertin IT[1] est un éditeur de solutions logicielles dédiées à la sécurité des systèmes d'information, au traitement avancé des données numériques et vocales, et à la Speech Intelligence (Bertin IT, 2017). Basée à Montigny, en région parisienne, l'entreprise Bertin IT regroupe près de 120 collaborateurs en France et à l'étranger.
Les diverses solutions proposées par l'entreprise sont regroupées sous des entités spécifiques. Les solutions dédiées à la sécurité des systèmes d'information sont regroupées sous l'entité CrossinG, qui assure l'isolation des systèmes d'information critiques et sécurise les échanges (Bertin IT, 2017). La Speech Intelligence est le sujet de la gamme Vecsys, proposant MediaSpeech pour le *speech-to-text* multilingue et MobileSpeech pour la commande vocale. Enfin, les solutions consacrées au traitement avancé des données numériques s'articulent autour de MediaCentric et de la plateforme AMI Enterprise Intelligence[2] (Ami EI). MediaCentric offre à ses utilisateurs des capacités d'acquisition multi-sources en continu et analyse en profondeur des contenus multimédias et multilingues. Quant à Ami EI, il est l'outil de veille stratégique de Bertin IT, développé et maintenu par l'équipe R&D basée à Montpellier. L'équipe travaillait auparavant pour AMI Software, racheté par le groupe CNIM en mai 2015 et intégré à Bertin IT afin de se positionner comme un acteur majeur dans le domaine de la veille numérique. L'outil AMI EI offre à ses utilisateurs

— une collecte ciblée en continu sur des millions de sources Web multilingues ;

1. https://www.bertin-it.com/
2. https://www.bertin-it.com/intelligence-numerique/solution-veille-strategique-intelligence-competitive/

— l'indexation, le filtrage et le classement automatiques de l'ensemble de leurs données ;
— l'analyse sémantique et la détection de concepts émergents.

C'est au sein de l'équipe R&D, basée à Montpellier, que j'ai effectué mon stage. L'équipe de Montpellier se divise en 3 équipes : l'équipe Développement, qui s'occupe du développement *back* et *front* et des tests de l'outil AMI EI, l'équipe Sourcebox, chargée des réseaux et de l'infrastructure pour les sources de la veille, et l'équipe Support, regroupant toutes les actions de réponse d'aide pour l'utilisation des produits de l'outil. L'équipe se compose également d'une ingénieur chercheuse, Leila Khouas, qui a dirigé mon stage. Leila s'occupe du service Recherche, c'est-à-dire qu'elle encadre les éventuels doctorants et gère les différents projets de recherche de niveau national ou européen. Elle est également la référente sur tous les projets long terme et fouille de textes de l'équipe Développement. La principale fonctionnalité d'AMI EI est de permettre à ses utilisateurs de mettre en place des veilles appelées plans de collectes, composés de scénarios correspondant aux différents sujets de la veille en question. Ces veilles sont quotidiennes. Le veilleur fournit plusieurs mots ou expressions-clé de son choix, combinés (*and*) ou non (*or*). Il définit le nombre de documents maximum qu'il souhaite recevoir et choisit ou non d'éliminer les doublons. Les sources de la veille sont définies par le veilleur. Elles sont ensuite parcourues par Sourcebox, qui les met à disposition des collectes quotidiennes du veilleur. L'outil analyse ensuite les données collectées.

2 Présentation du sujet

La veille stratégique est la collecte permanente d'informations sur les avancées et les orientations stratégiques de la concurrence en matière de produits, de techniques de production, de modes de commercialisation ou de communication (Emarketing, 2020). La veille stratégique met donc à disposition de l'utilisateur un très grand nombre de données textuelles contenant autant d'informations présentées de manière non-structurée (Jean-Louis, 2011). Dans ce contexte, il est pertinent de pouvoir présenter ces informations sous une forme structurée, "en se focalisant sur celles jugées pertinentes vis-à-vis du domaine d'intérêt considéré" (Jean-Louis, 2011, p. 6), et de les analyser.

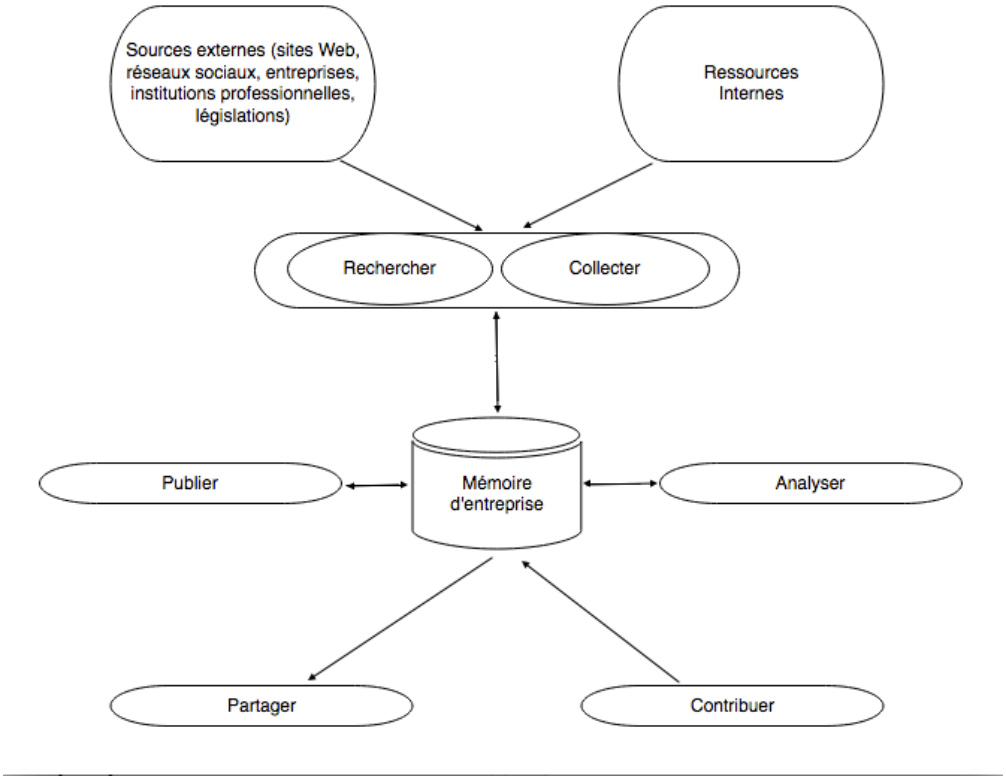


FIGURE 1.1 – Schéma du processus de veille

Bertin IT souhaite enrichir la solution Ami EI en y intégrant un outil d'extraction d'évènements afin de présenter au veilleur les évènements pertinents présents dans chacun des textes de la veille qu'il aura mise en place. Cet outil devrait idéalement être capable d'extraire des évènements en plusieurs langues et pouvoir s'adapter facilement à l'architecture déjà en place. L'objet de ce stage était d'offrir à l'entreprise un état de l'art fouillé sur le sujet, de présenter une exploration des solutions d'extraction d'évènements disponibles en Open Source ainsi qu'au niveau industriel et de mettre en place une maquette d'extraction d'évènements.

Chapitre 2

État de l'art

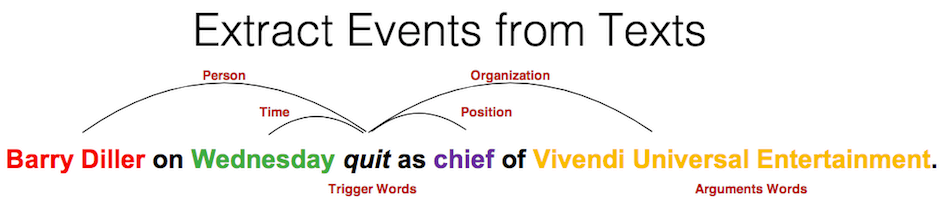
1 Définition de l'extraction d'évènements

L'extraction d'évènements est une application du Traitement Automatique des Langues (ci-après, TAL) et une branche de l'extraction d'information consistant à extraire de manière structurée des informations sur des évènements présents de manière non structurée dans des textes écrits.

Définir la tâche d'extraction d'évènements nécessite de définir la notion d'évènement en tant que tel. Or, au même titre que les entités nommées, le terme *évènement* est souvent défini de manière empirique par rapport à la tâche donnée sans qu'existe vraiment une définition standard (Arnulphy, 2012). Dans sa thèse de doctorat *Désignations nominales des évènements : Étude et extraction automatique dans les textes*, Béatrice Arnulphy consacre un long chapitre à cette problématique et explore la définition d'évènement dans différents domaines académiques. Partant de là, nous pouvons retenir trois notions essentielles propres à l'évènement : un évènement se **produit**, implique des **participants** et provoque **un changement d'état** ; il est ancré dans le **temps** (et implique donc une idée de fin, même si celle-ci n'est pas toujours réalisée) ; enfin, il a une **importance**, un **impact** dans notre monde. Ces trois notions se retrouvent d'ailleurs dans le guide d'annotation de la campagne d'évaluation ACE, l'*ACE English Annotation Guidelines for Events* (2005), dont les premiers mots sont *"An Event is a specific occurrence involving participants. An Event is something that happens. An Event can frequently be described as a change of state"* (ACE, 2005, p. 5), que l'on pourrait traduire par *"Un évènement est une occurrence spécifique impliquant des participants. Un évènement est quelque chose qui se produit. Un évènement peut très souvent être décrit comme un changement d'état"*. L'extraction d'évènements peut également se définir comme

une réponse à la question *"Who did What to Whom and Where and When"* (*Qui a fait à qui, où et quand ?*) (Netowl, 2019) et s'apparente donc à une tâche de remplissage de formulaire (Kodelja, Besancon & Ferret, 2017). Cet aspect de remplissage de formulaire, introduit lors de la seconde édition de la campagne d'évaluation *Message Understanding Conference*[1] (ci-après, MUC) en 1989, prédomine toujours à l'heure actuelle. Dans cette veine, le développement de guides d'annotation d'évènements va de pair avec les campagnes d'évaluation et classent les évènements par type, dont le nombre varie selon les campagnes et les guides. Chaque évènement dans son ensemble suppose de remplir un formulaire structuré en quatre parties contenant :

1. la mention de l'évènement et de son type ;
2. la mention du déclencheur (ci-après, trigger) de l'évènement, qui est souvent un verbe ou un substantif ;
3. le relevé des arguments de l'évènement ;
4. le relevé des rôles correspondant aux arguments.



Trigger	Quit (a "Personnel/End-Position" event)	
Arguments	Role = Person	Barry Diller
	Role = Organization	Vivendi Universal Entertainment
	Role = Position	Chief
	Role = Time-within	Wednesday (2003-03-04)

FIGURE 2.1 – Schéma d'extraction d'évènements et de remplissage de formulaire

Source : http ://tcci.ccf.org.cn/summit/2017/dlinfo/002.pdf

Pour ce faire, la tâche d'extraction d'évènements se divise en plusieurs sous-tâches généralement traitées de manière séquentielle (Kodelja, Besançon & Ferret, 2017). La distinction de ces sous-tâches s'opère au niveau du type d'extraction d'évènements. De manière générale, l'extraction d'évènements est traitée de manière *closed-domain* (ci-après, supervisée),

1. https://www-nlpir.nist.gov/related_projects/muc/

c'est-à-dire que le modèle d'extraction se base sur un ou plusieurs types d'évènements précis, définis de manière structurée en amont de l'extraction d'évènements à proprement parler. Chaque évènement est classé par type, dont le nombre varie selon les guides d'annotations. Nous nous basons ici sur l'*ACE English Annotation Guidelines for Events*[2], devenu un standard dans le domaine. Selon ce guide, dans la phrase *Yesterday, some demonstrators threw stones at soldiers in Israel* (Figure 2.2), l'outil d'extraction doit identifier l'évènement

FIGURE 2.2 – Exemple d'évènement de type *Attack*

Source : https ://www.aclweb.org/anthology/P17-1038.pdf

comme un évènement de type *Conflict/Attack* déclenché par le trigger *threw* puis extraire tous les arguments de l'évènement correspondant aux *argument-roles*. Un évènement de type *Conflict/Attack* suppose les arguments suivants :

- celui ou celle qui lance l'attaque (***attacker***) - ici, *demonstrators* ;
- celui ou celle qui est visé(e) par l'attaque (***target***) - ici, *soldiers* ;
- l'instrument utilisé dans l'attaque (***instrument***) - ici, *stones* ;
- le moment de l'attaque (***time***) - ici, *yesterday* ;
- le lieu de l'attaque (***place***) - ici, *Israel*.

L'extraction d'évènements supervisée suppose donc quatre tâches : repérer l'élément déclencheur, identifier le type d'évènement grâce à cet élément déclencheur, repérer dans le texte les arguments liés à l'évènement et identifier le rôle de ces arguments. Une cinquième tâche, la coréférence, se greffe aux quatre autres ; elle consiste à distinguer et relier les différentes mentions du même évènement.

De l'autre côté du spectre se trouve l'extraction d'évènements dite en *open domain* (ci-après, non-supervisée). Moins étudiée que sa consœur, elle suscite un regain d'intérêt depuis le début des années 2010 compte tenu de l'explosion du *big data* et des médias sociaux. L'extraction d'évènements non-supervisée consiste en "l'exploration de grandes masses de contenu textuel et l'acquisition automatique de structure d'évènements à partir de ces textes" (Besançon, 2016, p. 4). Elle n'est pas guidée par des schémas d'évènements préconçus (Liu, Huang & Zhang, 2019). Les tâches principales de ce type d'extraction d'évènements sont au nombre deux, à savoir la conception de schémas d'extraction et l'extraction en elle-même à partir de

2. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

ces schémas (Wei & Wang, 2019).
L'extraction d'évènements est une tâche applicable à de nombreux domaines tels que l'informatique (expansion des bases de connaissances), la défense & l'armée, la finance ou encore la médecine (Wei & Wang, 2019). Cela se reflète beaucoup dans les campagnes d'évaluation. La première campagne de la MUC, en 1987, avait pour objet l'extraction d'évènements issus de messages militaires[3], et le programme Jasper[4], dont le but était de fournir aux opérateurs financiers des informations financières en temps réel, fut créé au milieu des années 80 ("Information extraction", 2020). L'extraction d'évènements est donc étudiée depuis près de trois décennies mais reste une tâche complexe, dont les modèles sont à l'heure actuelle difficilement généralisables et applicables à plusieurs langues, et ce malgré qu'elle bénéficie des avancées spectaculaires dans de nombreuses tâches connexes du TAL dont elle dépend (l'analyse morpho-syntaxiques, la reconnaissance d'entités nommées, etc) (Wei & Wang, 2019).

2 Approches en extraction d'évènements

Au même titre que d'autres tâches typiques du TAL, comme l'analyse syntaxique ou la reconnaissance d'entités nommées, l'évolution des méthodes en extraction d'évènements va de pair avec les tendances en cours et les évolutions technologiques. Ce chapitre est consacré aux méthodes utilisées en extraction d'évènements depuis la première édition de la MUC (voir *2.3. Campagnes d'évaluation*) en 1987, et est construit chronologiquement, traitant d'abord des modèles dits de *pattern-matching* (ci-après, reconnaissance de motifs) en passant par les méthodes issues de l'apprentissage automatique puis des réseaux de neurones profonds.

2.1 Approches basées sur la reconnaissance de motifs

Les toutes premières approches de l'extraction d'évènements reposaient sur des systèmes à base de règles et d'expressions régulières définies manuellement (Kodelja, Besancon & Ferret, 2017). Compte tenu de la lourdeur de la tâche et du manque d'adaptabilité, ces systèmes ont rapidement été supplantés par des systèmes basés sur des méthodes d'extraction de motifs, où l'annotation d'un corpus remplaçait la conception des règles (Kodelja, Besancon & Ferret, 2017). La correspondance de motifs suppose qu'à un trigger X correspond une ou plusieurs suite d'entités (des **motifs**) qui, selon leur placement dans la phrase, correspondent

3. https://en.wikipedia.org/wiki/Message_Understanding_Conference
4. https://www.aclweb.org/anthology/A92-1024.pdf

aux arguments de l'évènement enclenché par le trigger X. Par exemple, un grand nombre de phrases contenant un évènement de type *kidnapping* seront à la voix passive ; le sujet du verbe sera alors la victime (la personne qui a été enlevée) et le complément d'agent, le kidnappeur (la personne qui procède au kidnappping). Partant, un trigger tel que *to kidnap* engendrera un motif de type ***PER_kidnap_PER*** qui, appliqué à un texte enrichi d'analyses morpho-syntaxique et syntaxique, sera capable d'extraire les arguments des évènements de types *kidnapping* dans les phrases dont le verbe est le verb *to kidnap* par la correspondance entre le motif générique et la structure de la phrase analysée.

2.2 Approches basées sur l'apprentissage automatique

Les méthodes d'extraction d'évènements basées sur l'apprentissage automatique considère la tâche d'extraction d'évènements comme "une tâche de classification de séquences" (Kodelja, Besancon & Ferret, 2017, p. 5). Ces méthodes reposent généralement sur des algorithmes de classification classiques en apprentissage automatique, tels que le SVM-*Support Vector Machine* ou le ME-*Maximum Entropy* (Wei & Wang, 2019). L'extraction d'évènements par l'apprentissage automatique relève de l'apprentissage supervisé et nécessite donc un corpus d'entraînement annoté précisément. Le but de ces méthodes est d'extraire à partir de textes bruts différents traits (lexicaux, syntaxiques, sémantiques) représentés par des vecteurs, et utiliser ces traits, ainsi que les étiquettes leur correspondant dans les corpus d'entraînements, pour entraîner les classifieurs. Un texte étant un ensemble de phrases traitées comme des séquences de *tokens* (Kodelja, Besancon & Ferret, 2017), les classifieurs doivent alors déterminer si le *token t* de la phrase correspond à un trigger ou à un argument, et, partant, indiquer le type d'évènement ou le rôle de cet argument. Les méthodes varient selon de nombreux facteurs, notamment l'architecture du modèle (architecture séquentielle ou architecture conjointe), le nombre de classifieurs ou le type de traits utilisés pour l'entraînement. Puisque l'extraction d'évènements est par définition envisagée de manière séquentielle, les architectures séquentielles ont d'abord été plus systématiques. Les architectures séquentielles traitent les différentes tâches d'extraction d'évènements de manière séquentielles et extraient d'abord le trigger puis ses arguments. En 2013, Qi Li *et al.* furent parmi les premiers à proposer une méthode conjointe afin de prédire conjointement les déclencheurs et leurs arguments (Li, Ji & Huang, 2013) et contrer l'un des problèmes inhérents aux méthodes en architecture séquentielle, à savoir "la propagation d'erreurs, puisque les déclencheurs et les arguments sont prédits de manière isolée par des classifieurs indépendants" (Li, Ji & Huang, 2013, p. 73). Dans sa thèse *Neural Methods for Event Extraction* consacrée à l'amélioration de la per-

formance en extraction d'évènements, Emanuela Boros[5] utilise les résultats présentés dans l'article système de Qi li *et al.* en tant que *baseline*, c'est-à-dire en tant que mesure de comparaison. Les méthodes d'apprentissage automatique ont réduit l'intervention humaine dans l'extraction d'évènements mais en terme d'efforts, "l'effort d'élaboration de règles [a été remplacé] par un effort d'ingénierie des représentations aussi conséquent" (Kodelja, Besancon & Ferret, 2017, p. 5).

2.3 Approches basées sur l'apprentissage profond

L'apprentissage profond est une branche de l'apprentissage automatique dont le principe repose sur l'utilisation d'algorithmes à plusieurs couches (d'où le terme *profond*), principalement des réseaux de neurones (Bengio, 2019). Depuis quelques années, les réseaux de neurones sont revenus en force sur le devant de la scène de l'intelligence artificielle grâce à l'amélioration de la puissance de calcul des ordinateurs et à l'accès à un très grand nombre de données étiquetées (Bengio, 2019). L'apprentissage profond s'est propagé à de nombreuses tâches du TAL, par exemple la reconnaissance d'entités nommées (Wei & Wang, 2019), et n'a pas échappé au domaine de l'extraction d'évènements. En effet, les réseaux de neurones peuvent pallier les inconvénients des modèles cités ci-dessus, à savoir le manque de généralisation, la complexité des représentations dans les modèles d'apprentissage machine, la propagation d'erreurs ainsi que l'effort humain considérable en annotation et donc, le faible nombre de corpus annotés de manière précise.

2.4 Autres approches

L'écart entre la complexité de certaines méthodes issues de l'intelligence artificielle et les résultats engendrés poussent à explorer d'autres méthodes se basant sur des outils de nature plus linguistiques, par exemple FrameNet[6] ou le *Semantic Role Labeling*[7] (ci-après, SRL). FrameNet est une base de données lexicale reposant sur la théorie de la *Frame Semantics* développée par Charles J. Fillmore selon laquelle "la signification d'un mot ne peut être comprise sans avoir accès à toutes les connaissances qui se rapportent à ce mot" (ELLO, 2020). Cette base de données classe les éléments textuels par cadre sémantique (Framenet, 2020) et contient plus de 1 200 cadres sémantiques, 13 000 unités lexicales (appariement d'un

5. http://www.theses.fr/2018SACLS302
6. https://framenet.icsi.berkeley.edu/fndrupal/
7. https://web.stanford.edu/~jurafsky/slp3/20.pdf

mot avec un sens) et 202 000 phrases d'exemple. Chaque cadre sémantique est composé d'arguments indispensables (*core*) et facultatifs (*non-core*). Ainsi, le verbe *to buy* est une action du cadre *Commerce* et est défini comme suit : *The Buyer wants the Goods and offers Money to a Seller in exchange for them* (FrameNet, 2020), où *Buyer* et *Goods* sont des arguments indispensables , tandis que *Money* et *Seller* sont des arguments non-indispensable . L'unité lexicale *buy.v* est quant à elle définie comme *COD : obtain in exchange for payment.*
Le SRL consiste à "analyser les propositions exprimées par certains verbes cibles de la phrase et à reconnaître, pour chaque verbe cible, tous les éléments de la phrase qui remplissent un rôle sémantique du verbe" (Carreras & Marquez, 2005, p. 152). Parmi les arguments sémantiques typiques on retrouve par exemple l'agent, le patient ou encore l'instrument (Carreras & Marquez, 2005). La structuration de Framenet et du *Semantic Role Labeling* présente des similitudes avec la structuration en formulaire dans l'extraction d'évènements. Or, si certains articles mentionnent des méthodes utilisant FrameNet ou le SRL pour l'extraction d'évènements, ceux-ci ne suscitent pas l'enthousiasme auquel on s'attendrait compte tenu de cette similitude de structure. Dans sa thèse[8] Emanuela Boros avance que, dans la situation [où le SRL pourrait permettre aux méthodes d'extraction d'évènements d'offrir de meilleurs résultats], "[il faut] faire face à l'effet *pipeline* : les erreurs dans les résultats du SRL, qui sont encore très élevées dans les outils actuels, sont susceptibles de nuire à l'extraction d'évènements, ce qui explique certainement pourquoi les outils du SRL ne sont pas largement utilisés pour l'extraction d'évènements" (Boros, 2015, p. 16).

3 Campagnes d'évaluation

La recherche en extraction d'évènements est jalonnée de nombreuses campagnes d'évaluation, dont certaines ont eu une très forte influence. Ces campagnes se déroulent très souvent en plusieurs éditions et les tâches s'enrichissent et se complexifient au fil de celles-ci. Les campagnes d'évaluation les plus célèbres, et dont l'influence s'est fait le plus ressentir, ont toutes été financées par des organismes américains. Dans un souci d'exhaustivité, nous présenterons également les campagnes les plus importantes d'extraction d'évènements dans le domaine biomédical, ainsi que certaines campagnes d'évaluation financées par l'Union européenne.
Comme indiqué ci-avant, la première d'entre elles, la MUC, débuta en 1987 et courut sur

8. http://www.theses.fr/2018SACLS302

sept éditions en dix ans. Financée par la DARPA[9] (Defense Advanced Research Projects Agency), son but premier était d'évaluer et d'encourager la recherche sur l'analyse automatisée de messages militaires contenant des informations textuelles (Grishman & Sundheim, 1996). À partir de ces messages, les systèmes de chacun des participants devaient parvenir à remplir un formulaire (au format libre lors de la première édition, puis par la suite imposé par les organisateurs) dont les champs correspondaient aux informations tirées du texte et étaient par exemple la cause, l'agent, le moment et le lieu d'un évènement, ses conséquences, etc. De conférence en conférence, le nombre de champs a augmenté (Message Understanding Conference, 2020). Quant aux thèmes abordés et à leurs sources, ils ont évolué au fil du temps ; lors des deux premières éditions, les textes à analyser étaient des messages militaires, qui sont devenus des coupures de presse traitant d'activités terroristes en Amérique latine lors des deux suivantes, puis des coupures de presse sur des thèmes d'ordre plus civil (notamment le rachat d'entreprise) lors des trois dernières éditions (Grishman & Sundheim, 1996).

9. https://www.darpa.mil/

TST1-MUC3-0080

BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) – [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS HE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND INTO A BLUE RENAULT.

HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RELEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT.

LAST WEEK, FEDERICO ESTRADA VELEZ HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

0. MESSAGE ID	TST1-MUC3-0080
1. TEMPLATE ID	1
2. DATE OF INCIDENT	03 APR 90
3. TYPE OF INCIDENT	KIDNAPPING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"THREE HEAVILY ARMED MEN"
6. PERPETRATOR: ID OF ORG(S)	"THE EXTRADITABLES"
7. PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: "THE EXTRADITABLES"
8. PHYSICAL TARGET: ID(S)	*
9. PHYSICAL TARGET: TOTAL NUM	*
10. PHYSICAL TARGET: TYPE(S)	*
11. HUMAN TARGET: ID(S)	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR")
12. HUMAN TARGET: TOTAL NUM	1
13. HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL: "FEDERICO ESTRADA VELEZ"
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPE(S)	*
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17. EFFECT ON PHYSICAL TARGET(S)	*
18. EFFECT ON HUMAN TARGET(S)	-

FIGURE 2.3 – MUC-3 - Exemple de message et son formulaire associé

Source : https ://www.aclweb.org/anthology/X96-1047.pdf

La vision de l'extraction d'évènements en tant que tâche de remplissage de formulaire, l'introduction des mesures de précision et de rappel dès la deuxième édition, et l'évaluation des entités nommées et de la coréférence dès la sixième édition ont fait que les campagnes d'évaluation MUC ont en grande partie façonné les programmes de recherche en matière d'extraction d'évènements (Grishman & Sundheim, 1996).

Succédant à la MUC, la campagne Automatic Content Extraction [10] (ci-après, ACE), financée par la NIST [11] (National Institute of Standards and Technology), débuta en 1999 et courut sur 6 éditions. Son but était de détecter, à partir de textes en anglais, mandarin et arabe, les entités nommées et les relations entre ces entités puis, à partir de l'édition de 2005, les évènements (Doddington & al, 2004). Dans les campagnes ACE, "l'objectif était d'atteindre un certain degré de généralité dans les évènements détectés grâce à un inventaire d'évène-

10. https://www.ldc.upenn.edu/collaborations/past-projects/ace
11. https://www.nist.gov/

ments élémentaires" (Grishman, 2010, p. 2928). Par conséquent, le guide d'annotation *ACE English Annotation Guidelines for Events*, développé dans le cadre de la campagne ACE 2005, décrit huit types d'évènements assez généralistes, dont les thèmes parcourent un large spectre, allant du divorce aux fusions/acquisitions en passant par le changement de direction dans les entreprises (voir *2.4.1. Corpus ACE 2005*). Ce guide d'annotation est depuis lors devenu un standard dans l'annotation d'évènements, et le corpus qui en a découlé constitue encore aujourd'hui le principal corpus d'entraînement des modèles d'extraction d'évènement (voir *2.4. Ressource disponibles*).

Dans les années 2010, la campagne *Text Analysis Conference* (ci-après, TAC) a succédé à la campagne ACE. Toujours financée par la NIST, les campagnes d'évaluation se concentraient surtout sur l'enrichissement de bases de connaissance. Seule la campagne de 2016 a consacré l'une de ses tâches à l'extraction d'évènements à proprement parler. Cependant, les participants étaient appelés à utiliser les corpus issus de campagnes précédentes (dont ACE) et aucun nouveau corpus n'a été fourni (TAC KBP 2016 Event Track, 2017). Depuis 2012 existe également la campagne DEFT (Deep Exploration and Filtering of Text)[12], financée par la DARPA, dont le but est l'analyse de données textuelles à très grande échelle via des réseaux de neurones profonds (Harris, 2014). Ce programme a permis l'émergence d'un nouveau standard d'annotation, le standard Entities, Relations, Events (ERE)[13], défini comme une version simplifiée du standard ACE 2015 mentionné ci-dessus, afin de fournir plus rapidement des données annotées. Bien que des corpus annotés selon ce standard aient servi lors de la campagne d'évaluation TAC, ils n'ont pas été publiés officiellement.

Parallèlement à ces campagnes, le domaine biomédical a également organisé de nombreuses campagnes d'évaluation en extraction d'évènements et fourni des corpus et des ressources dédiées (Biomedical text mining, 2020). En effet, le domaine biomédical suppose une grande masse de données textuelles, issue notamment de la littérature scientifique ou des dossiers des patients, et la détection des relations sémantiques et d'évènements sont des tâches centrales de la fouille de textes dans ce domaine (Bjorne & Salakoski, 2018). Parmi les campagnes dans ce domaine, nous pouvons retenir la série de campagnes BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology). Lancée conjointement par la Mitre Corporation[14], le CNIO[15] (Spanish National Cancer Research Center) et le NIH[16] (Na-

12. https://www.darpa.mil/program/deep-exploration-and-filtering-of-text
13. https://tac.nist.gov/2016/KBP/guidelines/summary_rich_ere_v4.2.pdf
14. https://www.mitre.org/
15. https://www.cnio.es/
16. https ://www.nih.gov/

tional Institutes of Health), cette série avait pour but d'évaluer les applications de fouille de textes dans le domaine médical et, partant, de créer un langage standard dans le domaine (BioCreAtIvE challenge, 2006). Depuis 2004, huit éditions on été organisées. La tâche d'extraction d'évènement de l'édition 2017 a permis la création du Biological Expression Language (BEL), représentant les observations biologiques de manière formelle et devenu un standard dans le domaine (Hayes, 2018).

Au sein de l'Union Européenne, nous pouvons retenir deux campagnes d'évaluation. La première, intitulée POLCON[17] (Political Conflict in Europe in the Shadow of the Great Recession) et lancée par l'EUI[18] (European University Institute) et l'ERC[19] (European Research Council), avait pour sujet l'extraction d'évènements concernant la protestation publique et l'identification croisée de mentions d'évènements, de chaînes et de réseaux d'évènements de référencement commun (POLCON, 2018). La seconde, organisée conjointement en 2019 par l'université de Koç et l'ERC dans le cadre du projet *The New Politics of Welfare : Towards an "Emerging Markets" Welfare State Regime*[20] avait pour but d'identifier les nouveaux régimes à tendances sociales (*welfare regimes*) au sein des pays émergents et d'expliquer leur apparition.

4 Ressources disponibles

Les campagnes d'évaluation décrites ci-avant fournissent systématiquement aux participants un corpus annoté selon les guides d'annotation mis en place pour les dites campagnes. Il existe deux écoles d'annotations d'évènements (Ahn, 2006) : l'annotation dite temporelle et l'annotation dite complexe. L'annotation temporelle considère un évènement "comme un mot indiquant un nœud dans un réseau de relations temporelles" (Ahn, 2006, p. 1) tandis que l'annotation complexe considère un évènement "comme une structure complexe, reliant des arguments qui sont eux-mêmes des structures complexes, mais avec seulement une information temporelle auxiliaire" (Ahn, 2006, p. 1). La définition de l'extraction d'évènements présentée ci-dessus envisage l'annotation de type complexe, et le formulaire représentant l'évènement, son trigger et ses arguments fait écho par sa structure à cette complexité. Par conséquent, toutes les ressources issues de campagnes d'évaluation ne sont pas équivalentes, et toutes ne sont pas utilisables si l'on envisage l'extraction d'évènements en tant que rem-

17. https://www.eui.eu/Projects/POLCON
18. https://www.eui.eu/
19. https://erc.europa.eu/
20. https://cordis.europa.eu/project/id/714868/fr

plissage de formulaire. Les corpus Temp_eval_2[21] et Temp_Eval_3[22], par exemple, issus des tâches éponymes de la campagne SEmEval sont fournis pour "évaluer les évènements, les expressions temporelles et les relations temporelles" (SemEval-2, 2008) mais sont annotés de manière temporelle, sans la dimension complexe (notamment en terme sémantique) indispensable pour la tâche d'extraction d'évènements telle que décrite ci-dessus. Dans ces deux corpus, les évènements sont répartis en six classes différentes : la classe *Reporting*, la classe *Perception*, la classe *Aspectual*, la classe *I_action*, la classe*I_state*, la classe *State* et la classe *Occurrence*. Comme on peut le voir sur l'exemple suivant, extrait du corpus Temp-Eval-2 (Figure 2.4),

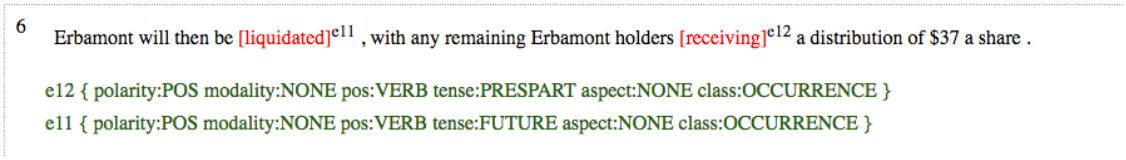


FIGURE 2.4 – Extrait du corpus Temp-Eval-2

Source : http ://semeval2.fbk.eu/semeval2.php ?location=data

la phrase *"Erbamont will then be liquidated , with any remaining Erbamont holders receiving a distribution of $37 a share"* contient deux évènements de la même classe *Occurrence* bien qu'ils soient sémantiquement très éloignés et supposeraient deux évènements de deux types distincts, remplissant deux formulaires distincts. Ces deux corpus ne peuvent donc constituer une ressource solide pour l'extraction d'évènements. Il en va de même pour le *NewsReader MEANTIME corpus*, issu de *NewsReader*, qui est un projet financé par le CORDIS[23] (acronyme de *Community Research and Development Information Service*), le service d'information sur la recherche et le développement de l'Union européenne. Ce projet vise à "mettre en place des bases de données d'évènements structurés issus de grands volumes de données financières et économiques afin d'améliorer la prise de décision" (CORDIS, 2017) et à donc extraire des évènements dans des textes. Or, le corpus annoté fourni en marge du projet annote les évènements selon trois classes seulement, la classe *grammatical*, la classe *speech-cognitive* et la classe *other*, ce qui en fait une ressource inutilisable pour les mêmes raisons qu'expliquées ci-dessus. Finalement, les corpus annotés conformément à la tâche décrite dans la section *Définition* sont uniquement les corpus des campagnes MUC et ACE.

21. http://semeval2.fbk.eu/semeval2.php?location=tasks
22. https://www.cs.york.ac.uk/semeval-2013/task1/
23. https://cordis.europa.eu/fr

Nous nous nous concentrerons seulement ici sur le corpus ACE 2005 car il est devenu un standard dans le domaine et est toujours systématiquement utilisé comme source première de données dans les systèmes d'extraction d'évènements supervisés, même les plus récents.

4.1 Corpus ACE 2005

Disponible en trois langues (anglais, mandarin et arabe), le corpus ACE 2005 est issu de la campagne du même nom. Le but des campagnes ACE est d'identifier dans des textes les **entités** (personnes, organisations, lieux, installations, armes, véhicules, entités géopolitiques), les **relations entre ces entités** (rôles, parties, localisation, proximité, relations sociales), les **évènements mentionnés** (interaction, mouvement, transfert, création, destruction) et les **valeurs** (*values*), catégorie regroupant les données numériques (pourcentage,...),les informations de contact (adresse,...), les données temporelles, les professions, les crimes, délits et les peines de prison (ces derniers pouvant effectivement servir d'arguments pour les évènements issues de la catégorie *Justice*). Les textes du corpus ACE 2005 sont issus de différents types de textes, notamment des articles de presse, des forums de discussion ou encore de conversations téléphoniques. Les évènements sont regroupés à travers huit types et 33 sous-types, comme le montre la table ci-dessous (Table 2.1).
Les guides d'annotation utilisés par les annotateurs sont mis à disposition des participants et sont au nombre de quatre, correspondant aux éléments ci-dessus :

- entités - ACE English Annotation Guidelines for Entities
- relations - ACE English Annotation Guidelines for Relations
- évènements - ACE English Annotation Guidelines for Events
- valeurs - ACE English Annotation Guidelines for Values & TIDES 2005 Standard for the Annotation of Temporal Expressions

Chaque phrase du corpus peut contenir zéro, un ou plusieurs évènements, et chaque évènement suppose la participation d'une ou plusieurs entités, à laquelle est assigné un rôle. Ce couple entité-rôle correspond aux arguments de l'évènement. Notons que certains évènements n'ont pas d'arguments.

5 Conclusion

En TAL, l'extraction d'évènements dans les textes est la forme la plus complexe des processus d'extraction d'information (Boros, 2015). Bien qu'un sujet de recherche dès les années

Types	Subtype
Life	Be-Born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-Ownership, Transfer-Money
Business	Start-Org, Merge-Org, Declare-Bankruptcy, End-Org
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
Personnel	Start-Position, End-Position, Nominate, Elect
Justice	Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

TABLE 2.1 – Tableau des types & sous-types d'évènements répertoriés pour ACE 2005

Source : Automatic Content Extraction English Annotation Guidelines for Events (2005)

90, l'extraction d'évènements n'a pas encore d'outil stable en Open Source qui lui est propre, à l'instar, par exemple, des bibliothèques Stanza (Python) ou CoreNLP (Java) de Stanford NLP pour les entités nommées. La recherche dans ce domaine s'est très longtemps concentrée sur l'extraction thématique, ce qui a eu des conséquences qui se font encore ressentir aujourd'hui. D'abord, l'extraction d'évènements thématique, et donc supervisée, suppose presque toujours la mise à disposition de données annotées, extrêmement coûteuses à mettre en place et peu adaptable à différentes langues. Ces ressources sont de très grande qualité mais malheureusement insuffisantes, ce qui présente des désavantages à la fois pour les systèmes d'apprentissage automatique et les systèmes d'apprentissage profond. Pour les premiers, ces données annotées très précisément supposent un effort considérable dans l'élaboration des représentations ; pour les seconds, elles ne sont pas suffisantes, notamment pour la détection et l'extraction des arguments, car la précision des annotations est telle qu'elle supposerait un corpus d'entraînement beaucoup plus fourni. En outre, 33 types d'évènements sont représentés dans le corpus ACE 2005 mais de nombreuses expressions de ces types d'évènements ne sont pas incluses, ce qui limite les performances (Cao *et al.*, 2015). A contrario, certains types d'évènements sont sur-représentés par rapport à d'autres, notamment au niveau des

triggers (Table 2.2), ce qui n'est pas bon pour l'apprentissage (Boros, 2015).

Trigger Types (frequency)	
Conflict.Attack (1252)	Movement.Transport (607)
Life.Die (488)	Personnel.End-Position (196)
Contact.Meet (190)	Personnel.Elect (143)
Transaction.Transfer-Money (140)	Life.Injure (116)
Justice.Charge-Indict (107)	Contact.Phone-Write (107)
Transaction.Transfer-Ownership (101)	Justice.Trial-Hearing (100)
Justice.Sentence (94)	Personnel.Start-Position (92)
Justice.Arrest-Jail (88)	Justice.Convict (75)
Conflict.Demonstrate (72)	Life.Marry (58)
Justice.Sue (55)	Life.Be-Born (46)
Business.Declare-Bankruptcy (40)	Justice.Appeal (39)
Business.Start-Org (38)	Justice.Release-Parole (35)
Business.End-Org (33)	Life.Divorce (28)
Justice.Fine (28)	Justice.Execute (20)
Business.Merge-Org (18)	Personnel.Nominate (11)
Justice.Acquit (7)	Justice.Extradite (3)
Justice.Pardon (2)	

TABLE 2.2 – Nombre de triggers par types d'évènements dans le corpus ACE 2005

Source : https ://www.aclweb.org/anthology/C16-1114.pdf

La raison pour laquelle le corpus ACE est devenu aujourd'hui un standard est peut-être d'ordre chronologique ; la dernière campagne ACE était la campagne de 2005 et les campagnes TAC, qui lui ont succédé, se sont davantage concentrées sur l'expansion de bases de connaissance que sur l'extraction d'évènements proprement dite. Ensuite, le début des années 2010 a vu l'avènement du *big data* et beaucoup ont probablement pensé que l'accès à cette énorme masse de données allait améliorer les performances. Ce phénomène a surtout orienté la recherche vers l'extraction d'évènements non-supervisée, notamment sur Twitter, mais celle-ci est encore loin, très loin d'être mature à l'heure actuelle. Pour l'extraction d'évènements supervisée, d'autres enjeux, et non des moindres, sont nés du *big data*, notamment la problématique de la normalisation des données textuelles issues des réseaux sociaux[24]. Troisièmement, du point de vue industriel, l'extraction d'évènements reste, à notre connaissance, une niche. Le seul outil d'extraction d'évènements à proprement parler, c'est-à-dire qui mentionne exactement le terme d'extraction d'évènements, disponible sur le marché est,

24. Les textes issus des réseaux sociaux peuvent être bruités et contenir des tournures ou de la syntaxe s'éloignant de la norme, ce qui les rend plus difficiles à analyser

à notre connaissance, l'outil NetOwl Extractor[25], de l'entreprise NetOwl[26]. Cependant, aucune démonstration n'est disponible.

Les suites Natural Language[27] et AutoML Natural Language[28] de Google Cloud offre plusieurs fonctionnalités relatives au TAL comme la reconnaissance d'entités nommées, l'analyse syntaxique, l'analyse de sentiments ou la classification de textes. Les ressources proposées par ces suites peuvent sans doute aider les entreprises à envisager une solution d'extraction d'évènements mais cette solution n'est pas fournie clé en main et l'extraction d'évènements n'est pas une fonctionnalité spécifique. La suite Azure Cognitive Services[29] de Microsoft offre deux modules liés au TAL, le module Langage et le module Microsoft Speech. Le module Langage propose entre autres le service Analyse de textes[30] qui extrait d'un texte les mots-clés et les entités nommées, ainsi que le sentiment général, mais pas les évènements. L'outil Refinitiv Intelligent Tagging[31] (auparavant OpenCalais) offre la reconnaissance d'entités nommées, la reconnaissance des relations et le sujet traité.

À l'heure actuelle, l'extraction d'évènements demeure donc une tâche très complexe en TAL. Dans les prochaines années, si certains systèmes deviennent exploitables au niveau industriel, se posera toujours le défi des langues car les différents types de solutions sont généralement unilingues et difficiles à rendre multilingues.

25. https://www.netowl.com/event-extraction
26. https://www.netowl.com/
27. https://cloud.google.com/natural-language?hl=fr
28. https://cloud.google.com/natural-language?hl=fr#how-automl-natural-language-works
29. https://azure.microsoft.com/fr-fr/services/cognitive-services/#features
30. https://azure.microsoft.com/fr-fr/services/cognitive-services/text-analytics/
31. https://permid.org/onecalaisViewer

Chapitre 3

Développement du sujet

Mon stage au sein de Bertin IT s'est déroulé en trois temps. La première partie du stage a consisté en la rédaction d'un état de l'art sur l'extraction d'évènements destiné à l'équipe R&D ainsi qu'à l'entreprise en général, afin de leur permettre d'avoir une première vision de l'extraction d'évènements. Ce travail a permis de réaliser une synthèse de l'état de l'art telle que présentée ci-dessus et a constitué un document de travail tout au long du stage. Afin de manipuler des jeux de données et outils disponibles, la seconde partie du stage s'est concentrée sur la recherche de outils d'extraction d'évènements disponibles en Open Source, recherche guidée par les articles consultés lors de rédaction de l'état de l'art. Cette recherche s'est conclue par la prise en main et le test de deux outils dont les détails sont disponibles ci-dessous. La troisième partie du stage a été consacrée à la mise en place d'une maquette d'extraction d'évènements reposant sur une méthode inspirée des méthodes de reconnaissance de motifs (voir *2.2.1 Approches en extraction d'évènements basées sur la reconnaissance de motifs*).

1 Présentation d'outils Open Source guidée par la bibliographie

La rédaction de l'état de l'art destiné à l'entreprise a donné lieu à la lecture d'un grand nombre d'articles scientifiques sur l'extraction d'évènements. La deuxième phase du stage a été consacrée à la recherche d'outils disponibles en Open Source issus des articles en question. Plusieurs critères établis par l'entreprise ont guidé cette recherche, notamment le type d'extraction d'évènements traité, les langues supportées et la qualité de l'outil en terme

de F-Mesure. Dans ma recherche, je me suis concentrée sur des outils actuels, issus d'articles récents, reposant donc majoritairement sur des réseaux de neurones. Nous avons finalement mis en place et testé deux outils, dont les détails sont présentés ci-dessous (*Test des outils*). Ces deux outils étaient les seuls relativement aisés à prendre en main. En effet, dans les faits, peu d'outils et systèmes disponibles en Open Source sont complets, accessibles et testables. Notre constat est que l'apprentissage de ces outils et leur paramétrage sont laborieux ; il est difficile de trouver les bons paramétrages pour répliquer les résultats présentés dans les articles. Nous avons malgré tout pu mener diverses expérimentations et tester deux outils, ce qui nous a permis de dégager un premier éventail de possibilités, avantages et désavantages du développement de ce type de systèmes en terme industriel.
Les deux outils en question sont présentés ci-dessous.

1.1 Jointly Multiple Events Extraction (JMEE)

L'outil JMEE (Jointly Multiple Events Extraction) (Liu, Luo & Huang, 2018) propose un système en architecture conjointe reposant sur le corpus ACE 2005. Plus particulièrement, il se concentre sur un défi en extraction d'évènements, à savoir les phrases contenant plusieurs évènements. Ce phénomène est récurrent dans le corpus ACE 2005, où plus d'un quart des phrases sont concernées (Liu, Luo & Huang, 2018), ainsi que dans les textes en général. L'idée de l'outil JMEE est d'introduire des arcs de raccourcis syntaxiques, permettant de "réduire les sauts nécessaires d'un trigger à un autre dans la même phrase" (Liu, Luo & Huang, 2018, p. 1248). L'architecture proposée (Figure 3.1) utilise la représentation vectorielle GloVe[1], développée en Open Source par Stanford NLP. L'outil JMEE est disponible sur GitHub[2].

1. https://github.com/stanfordnlp/GloVe
2. https://github.com/lx865712528/EMNLP2018-JMEE

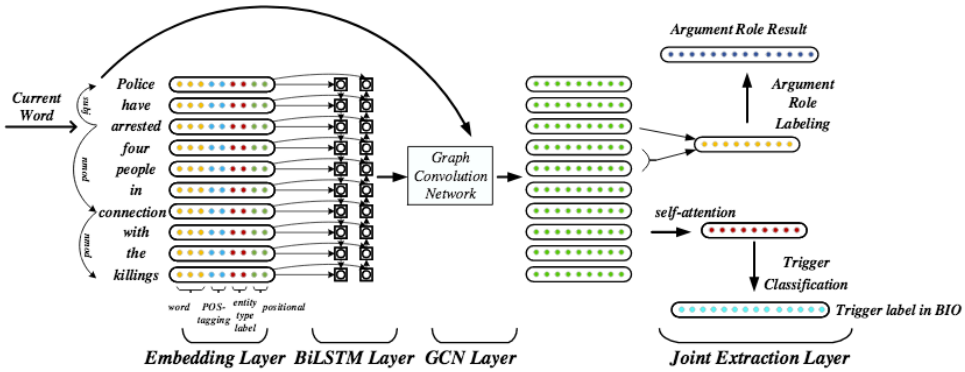


FIGURE 3.1 – Architecture JMEE

Source : https ://arxiv.org/pdf/1809.09078.pdf

1.2 Pytorch Solution of Event Extraction Task using BERT (Bert EE)

L'outil Bert EE[3] est une adaptation de l'outil JMEE. La principale modification se situe au niveau du modèle de langue. En effet, Bert EE utilise le modèle pré-entraîné de l'algorithme BERT de Google disponible sous PyTorch[4]. Bien que cette méthode soit une adaptation, nous avons décidé de la tester car l'algorithme BERT présente des résultats de pointe dans de nombreuses tâches de TAL (Horev, 2018). Le corpus d'entraînement de l'outil Bert EE est également le corpus ACE 2005. L'outil Bert EE est disponible sur GitHub[5].

1.3 Test des outils

Prétraitement du corpus

Ces deux outils nécessitent un prétraitement du corpus ACE 2005 pour enrichir celui-ci, notamment au niveau syntaxique. Ce prétraitement comprend la tokenisation, la lemmatisation, l'analyse morpho-syntaxique, l'analyse syntaxique et la représentation des dépendances. Ce prétraitement a été effectué à l'aide d'un script de prétraitement disponible sur GitHub[6].

3. https://github.com/nlpcl-lab/bert-event-extraction
4. https://github.com/huggingface/transformers
5. https://github.com/nlpcl-lab/bert-event-extraction
6. https://github.com/nlpcl-lab/ace2005-preprocessing

Ce script se base sur CoreNLP [7], l'une des librairies TAL de Stanford NLP. La sortie du script nous fournit l'entièreté du corpus en format JSON, enrichi des traits mentionnés ci-dessus, et découpé en corpus d'entraînement, de développement et de test.

Entraînement

Les méthodes JMEE et Bert EE sont des méthodes neuronales et reposent sur des modèles mathématiques complexes, dont les paramétrages sont difficiles. Nous avons entraîné les deux méthodes sur la composante d'entraînement du corpus ACE 2005. Compte tenu de sa volumétrie, les temps d'entraînement sont conséquents. Tout l'objet du test de ces deux méthodes a donc été de trouver le bon compromis entre obtenir des résultats satisfaisants et entraîner sur des temps raisonnables. Nous avons notamment fait plusieurs essais sur le nombre d'epochs [8]. Finalement, l'outil JMEE a été entraîné sur 10 epochs, et l'outil Bert EE, sur 3 epochs.

Résultats

Les résultats (Table 3.1 et Table 3.2) du test de ces outils Open Source (en terme de précision, rappel, F-Mesure) sont présentés selon les quatre critères suivants :

- l'identification du trigger correspond à l'identification du mot ou groupe de mots composant le trigger dans le texte ;
- la classification du trigger correspond à l'identification du type d'évènement que le trigger a déclenché ;
- l'identification de l'argument correspond à l'identification du mot ou groupe de mots composant l'argument dans le texte ;
- la classification de l'argument correspond à l'identification du rôle assigné à l'argument.

Ces résultats (Table 3.1 et Table 3.2) correspondent aux meilleurs indicateurs (précision, rappel, F-Mesure) obtenus en évaluant différents modèles JMEE et Bert EE construits selon différents paramétrages. L'évaluation a été réalisée en utilisant la composante test du corpus ACE 2005.

7. https://stanfordnlp.github.io/CoreNLP/
8. Une epoch est une itération complète sur l'ensemble des données d'apprentissage.

	Précision	Rappel	F-Mesure
Identification du trigger	0.54	0.40	0.46
Identification de(s) argument(s)	0.43	0.02	0.03

TABLE 3.1 – Présentation des résultats de l'outil JMEE

	Précision	Rappel	F-Mesure
Identification du trigger	0.59	0.74	0.66
Classification du trigger	0.52	0.65	0.58
Identification de(s) argument(s)	0.45	0.30	0.36
Classification de(s) argument(s)	0.43	0.29	0.34

TABLE 3.2 – Présentation des résultats de l'outil Bert EE

Conclusion

Lors de cette partie du stage, nous avons testé deux outils d'extraction d'évènements disponibles en Open Source et obtenu des résultats. Malgré cela, nous avons observé que de tels outils sont rares, difficiles à paramétrer et partant, difficilement testables. En outre, les résultats que nous avons obtenus étaient loin des résultats affichés par les auteurs.

2 Méthode mise en place

De notre point de vue, les outils d'extraction d'évènements disponibles en Open Source issus de méthodes neuronales présentent de nombreux désavantages pour la mise en place au niveau industriel. Ces méthodes constituent une "boîte noire difficilement accessible et modifiable" (Serrano *et al.*, 2013, p. 3). Les temps d'apprentissage sont très longs, le paramétrage est laborieux, et il est difficile de construire les bons modèles pour obtenir des résultats optimaux. À ce constat s'ajoute le fait que nous n'avons pas trouvé d'outil issu des méthodes de reconnaissance de motifs disponible en Open Source. Pour toutes ces raisons, nous avons décidé de mettre en place une maquette d'extraction d'évènements inspirée de méthodes de reconnaissance de motifs utilisant le corpus ACE 2005 (voir *2.4.1. Corpus ACE 2005*).

2.1 Méthodologie

Nous avons décidé d'adopter un processus itératif dans l'élaboration de notre maquette d'extraction d'évènements, c'est-à-dire d'effectuer des tests à chaque itération de la maquette afin d'avoir un repère d'évolution en terme de qualité (F-Mesure) et pouvoir progresser. Un script d'évaluation évaluant la sortie de notre maquette au niveau des triggers et des types d'évènements a donc été aussitôt mis en place. Notre corpus de référence était le corpus de test ACE 2005. Les idées pour améliorer la maquette à chacune des itérations ont été guidées par les articles de la bibliographie, et testées au fur et à mesure.

2.2 Évaluation

Le script d'évaluation évalue la sortie de notre maquette à chacune des itérations du processus. Les mesures utilisées pour cette évaluation sont les mesures préconisées en extraction d'évènements depuis la deuxième édition de la campagne MUC, à savoir la précision, le rappel et la F-Mesure. Les éléments évalués par notre script sont les éléments évalués au sein des systèmes d'extraction d'évènements, c'est-à-dire l'identification et la classification des triggers, ainsi que, dans un deuxième temps, l'identification et la classifications des arguments (voir *3.1.3 Test des outils*). Chaque évènement du corpus de référence est représenté par un formulaire stocké dans une structure de données de type hachage contenant le trigger et son type d'évènements, ainsi que ses arguments. Pour chaque phrase, le script d'évaluation compare donc les formulaires dits *golden* (Figure 3.2) issus du corpus de référence, et les formulaires issus de notre méthode. Si la phrase ne contient pas d'évènement, le formulaire est vide. Dans un premier temps, nous avons évalué uniquement l'identification et la

classification des triggers.


```
{
  "sentence": "Hariri submitted his resignation during a 10-minute meeting with the head
   of state at the Baabda presidential palace, outside the capital.",
  "golden-event-mentions": [
    {
      "trigger": {
        "text": "meeting",
        "start": 7,
        "end": 8
      },
      "arguments": [
        {
          "role": "Entity",
          "entity-type": "PER:Individual",
          "text": "the head of state",
          "start": 9,
          "end": 13
        },
        {
          "role": "Time-Holds",
          "entity-type": "TIM:time",
          "text": "10-minute",
          "start": 6,
          "end": 7
        },
        {
          "role": "Place",
          "entity-type": "FAC:Building-Grounds",
          "text": "the Baabda presidential palace, outside the capital",
          "start": 14,
          "end": 22
        },
        {
          "role": "Entity",
          "entity-type": "PER:Individual",
          "text": "Hariri",
          "start": 0,
          "end": 1
        }
      ],
      "event_type": "Contact:Meet"
    }
```


FIGURE 3.2 – Exemple de formulaire d'évènements *golden* issu du corpus de référence ACE 2005

La précision, le rappel et la F-Mesure de notre script d'évaluation ont été calculés comme suit :

$$precision = \frac{vraiPositifs}{vraiPositifs + fauxPositifs}$$

$$rappel = \frac{vraiPositifs}{vraiPositifs + fauxNegatifs}$$

$$F - Mesure = 2 \times \frac{precision \times rappel}{precision + rappel}$$

$$p = \frac{VP}{VP + FP}$$

$$r = \frac{VP}{VP + FN}$$

$$f1 = 2 \times \frac{p \times r}{p + r}$$

où

— les vrais positifs sont les éléments identifiés, correctement ;

— les faux positifs sont les éléments identifiés, incorrectement ;

— les vrais négatifs sont les éléments non-identifiés, correctement ;

— les faux négatifs sont les éléments non-identifiés, incorrectement.

Comme indiqué, ce script d'évaluation a été mis en place pour évaluer la sortie de notre maquette à chaque itération. Afin d'avoir une référence en terme de score, nous avons soumis la sortie de l'outil Bert EE à notre script d'évaluation (triggers et type d'évènement) pour pouvoir comparer les résultats issus de notre maquette à ceux issus de l'outil Bert EE. Les résultats de la sortie de l'outil Bert EE évalués par notre script sont les suivants (Table 3.3).

	Précision	Rappel	F-Mesure
Identification du trigger	0.49	0.79	0.60
Classification du trigger	0.46	0.75	0.57

TABLE 3.3 – Présentation des résultats de l'outil Bert EE évalués par notre script

2.3 Mise en oeuvre

Partant de l'hypothèse qu'à chaque type d'évènement correspond un ou plusieurs motifs d'entités (un agencement particulier d'étiquettes autour du trigger) très caractéristiques, le fait de déterminer de tels motifs permettrait alors d'identifier l'évènement, son type et éventuellement ses arguments. Notre approche consiste à tenter d'extraire ces motifs caractéristiques à chaque type d'évènements à partir du corpus d'entraînement ACE 2005, et d'en déduire ensuite un ensemble de règles permettant d'identifier le type d'évènements le plus plausible à partir des motifs observés sur un texte donné.
Compte tenu de sa nature itérative, l'élaboration de notre maquette d'extraction d'évènements s'est faite en plusieurs étapes. Dès le départ, nous avons à notre disposition deux éléments. Premièrement, le corpus ACE 2005 prétraité (voir *3.1.3 Prétraitement du corpus*) nous donne accès à l'étiquette morpho-syntaxique de chaque token du corpus. Ensuite, nous savons qu'en extraction d'évènements, l'évaluation des triggers comporte deux volets, l'identification (le *token*[9] du trigger) et la classification (son *type d'évènement* associé). Le trigger et son type d'évènements constituent donc le coeur de l'architecture de notre maquette, auxquels viendront se greffer les motifs au fur et à mesure des itérations. Notre maquette se construit donc sur l'association lemme[10] du trigger+étiquette morpho-syntaxique (ci après, trigger+étiquette morpho-syntaxique), stockée en tant que clé d'une structure de données de type hachage, dont les valeurs étaient le ou les types d'évènement associés au trigger. Cette structure de données constitue la base de la maquette et sera enrichie au fur et à mesure des itérations.

Prétraitement

Dans un contexte de veille, une maquette d'extraction d'évènements doit procéder à l'extraction à partir de textes bruts. Sachant que la clé de notre dictionnaire consiste en l'association du lemme du trigger et son étiquette morpho-syntaxique, nous devions, pour vérifier si la clé existait dans le texte, avoir à disposition le texte brut segmenté en tokens (ci-après, tokenisation) et analysé morpho-syntaxiquement.
Nous avons donc en premier lieu écrit un script de prétraitement similaire au script mentionné dans *3.1.3 Prétraitement du corpus* utilisable sur n'importe quel texte brut fourni en entrée dont la sortie est le texte prétraité phrase par phrase au format JSON. Outre la tokenisation,

9. "Nœud terminal" (*terminal node*), qui, du point de vue des traitements ultérieurs, ne sera pas découpé en unités plus petites (Webster & Kit, 1992)
10. Forme graphique choisie conventionnellement comme adresse dans un lexique (TLFI)

la lemmatisation et l'étiquetage morpho-syntaxique, ce script devait également fournir les entités nommées, nécessaires pour la construction des motifs d'entités, ainsi que leur indice (au niveau phrastique). Nous avons utilisé Stanza[11], la librairie TAL de Stanford NLP écrite en Python, pour sa facilité de prise en main. Stanza utilise les étiquettes morpho-syntaxiques du Penn Treebank Project[12] (Table 3.4).

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

TABLE 3.4 – Étiquettes morpho-syntaxiques du Penn Treebank Project

Source : https ://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

11. https://stanfordnlp.github.io/stanza/
12. https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Stanza obtient une très bonne F-Mesure en reconnaissance d'entités nommées : 88.8 sur 18 types (Qi, Zhang, Zhang, Bolton, Manning, 2020). Les différents types sont énumérés ci-dessous (Table 3.5).

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc.
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named langage
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	*first*, *second*, *third*, etc.
CARDINAL	Numerals that do not fall under another type

TABLE 3.5 – Types d'entités nommées reconnues par Stanza

Source : OntoNotes Release 5.0

Lors de l'évaluation, notre maquette d'extraction d'évènements a besoin de comparer des formulaires comparables. Or, certains éléments annotés dans le corpus de référence ne sont reconnus en tant qu'entités nommées ni par Stanza, ni par d'autres outils similaires. Nous les avons donc ajoutés nous-mêmes. Ces éléments sont :

— les numéros de téléphone ;
— les adresses e-mail ;
— les URL ;
— les crimes et délits.

Les numéros de téléphone, adresses mail et URL ont été récupérés à l'aide d'expressions régulières et annotés selon les étiquettes du corpus de référence (*Contact-Info :Phone-Number*, *Contact-Info :E-Mail*, *Contact-Info :URL*). Le besoin d'annoter des éléments du texte en tant que crimes et délits s'explique par le fait qu'ils peuvent constituer des arguments pour les évènements de type *Justice* (voir *2.4.1 Corpus ACE 2005*). Pour cette catégorie, et dans un souci d'optimisation, nous avons extrait de l'ensemble du corpus de référence la liste des entités annotées sous la catégorie *Crime* et l'avons enrichie à l'aide d'une liste générique[13]. Cette approche dite de *gazetteer*[14] nous a satisfaits car nous avons considéré les crimes et délits en tant que liste fermée.

Toujours dans l'optique de comparer des éléments comparables, les catégories d'entités du corpus de référence et de Stanza ont dû faire l'objet d'une harmonisation. Le corpus de référence relève 54 types d'entités différentes, contre 18 pour les outils de reconnaissances d'entités nommées les plus complets, dont Stanza. Pour la catégorie *ORG*, classique en reconnaissance d'entités nommées, le corpus de référence relève huit sous-types, tels que *ORG :Commercial* ou *ORG :Educational*. Pour la catégorie *PER*, le corpus de référence relève trois sous-types, tels *PER :Individual*. Pour ces deux exemples, nous avons adopté une approche généraliste, et modifié le corpus de référence. Cette décision a été guidée par une volonté de précision, c'est-à-dire pour ne pas assigner une mauvaise catégorie à une entité. Les autres entités ont pu être harmonisées ; par exemple, la catégorie *TIM :time* du corpus de référence comprend les entités reconnues en tant que *DATE* et *TIME* par Stanza.

Triggers et types d'évènements

Comme indiqué ci-dessus (*3.2.3. Mise en oeuvre*), notre base d'extraction est le couple trigger+étiquette morpho-syntaxique. Lors de la première itération, nous avons considéré

13. https://criminal.findlaw.com/criminal-charges/view-all-criminal-charges.html
14. https://en.wikipedia.org/wiki/Gazetteer

tous les triggers disponibles dans les corpus d'entraînement et de développement d'ACE 2005, sans modification aucune. Chacun des triggers est associé à un seul type d'évènement. Cette première itération "brute" nous permet d'obtenir un premier résultat et le positionner par rapport à l'évaluation de la sortie de BERT EE par notre script (Table 3.3). Les résultats de cette première itération sont présentés ci-dessous (Table 3.6).

	Précision	Rappel	F-Mesure
Identification du trigger	0.24	0.73	0.36
Classification du trigger	0.24	0.74	0.36

TABLE 3.6 – Résultats de la maquette - Première itération

Les résultats sont assez encourageants, étant donné que le dictionnaire (Figure 3.3) est sous forme brute et n'a pas été du tout retravaillé. On constate un rappel élevé et une précision plus faible.

```
{
bombing_NN :
    'Conflict:Attack',
appointment_NNS :
    'Contact:Meet'
former_JJ :
    'Personnel:End-Position'
    }
```

FIGURE 3.3 – Extrait du dictionnaire utilisé lors de la première itération de la maquette.

À la deuxième itération, nous avons procédé à trois modifications majeures au sein du dictionnaire. Premièrement, un trigger n'est plus associé à un seul type d'évènement mais à tous les types d'évènement possibles dans les corpus d'entraînement et de développement. Certains triggers peuvent en effet déclencher plusieurs types d'évènement. C'est le cas du verbe ***to kill***, qui peut générer un évènement de type *Life :Die* et un évènement de type *Conflict :Attack*. Ces différents types d'évènements possibles sont intégrés en tant que valeurs de la clé trigger+étiquette morpho-syntaxique dans notre dictionnaire à l'aide d'un tuple composé du type d'évènement et de sa fréquence d'apparition en pourcentage. Lors de la deuxième itération, si un motif trigger+étiquette morpho-syntaxique est présent dans une phrase, c'est désormais le type d'évènement le plus fréquent qui est assigné à ce trigger et le formulaire est créé.

Deuxièmement, nous avons procédé à une modification de la liste des triggers récupérée lors de la première itération. Ayant observé attentivement les fichiers de sortie de la première itération, nous avons remarqué que certains triggers, trop généralistes, entraînent une "sur-reconnaissance". Nous avons systématiquement supprimé les triggers issues des catégories morpho-syntaxiques suivantes :

- nombres cardinaux (CD) ;
- déterminants (DT) ;
- préposition & conjonction de subordination (IN) ;
- verbes modaux[15] (MD) ;
- pronoms personnels (PRP) ;
- pronoms possessifs (PRP$) ;
- adverbes (RB & WRB) ;
- pronoms relatifs (WP).

Nous avons observé les triggers trop généralistes. Nous avons systématiquement supprimé les triggers verbaux "to be", "to have" et "to get". Les triggers verbaux "to go", "to pay" et "to take" ont été supprimés pour l'élaboration de cette maquette mais pourront être réintroduits lorsqu'elle sera enrichie, notamment au niveau de l'analyse du contexte.

Troisièmement, nous avons adapté le script pour inclure les triggers enclenchés par des verbes à particule (*phrasal verbs*). Les verbes à particule sont la combinaison d'un verbe et d'une ou plusieurs particules, par exemple une préposition ou un verbe, qui en modifient son sens (Calvo Rigual, 2013). Ainsi, si le trigger est un trigger verbal à particule, il est conservé. Par exemple, bien que "to take" ait été supprimé, "to take over" est conservé car il enclenche ex-

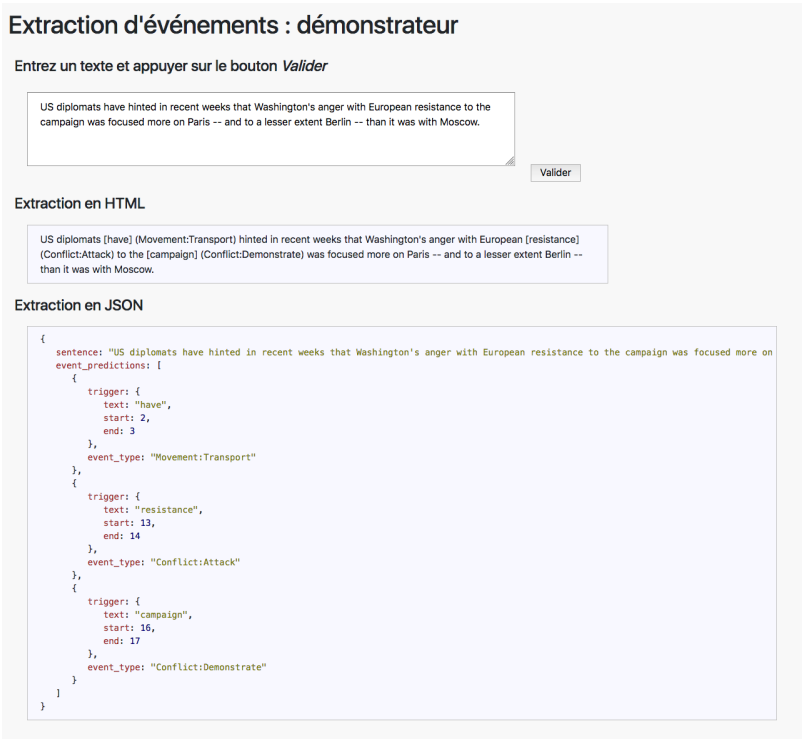
15. Dans le sens anglais du terme, c'est-à-dire *can*, *may*, *must*, *shall*, *will*, *could*, *might*, *should*, *would*.

plicitement un évènement de type *Personnel :Start-Position*. Les résultats de cette deuxième itération sont présentés ci-dessous (Table 3.7).

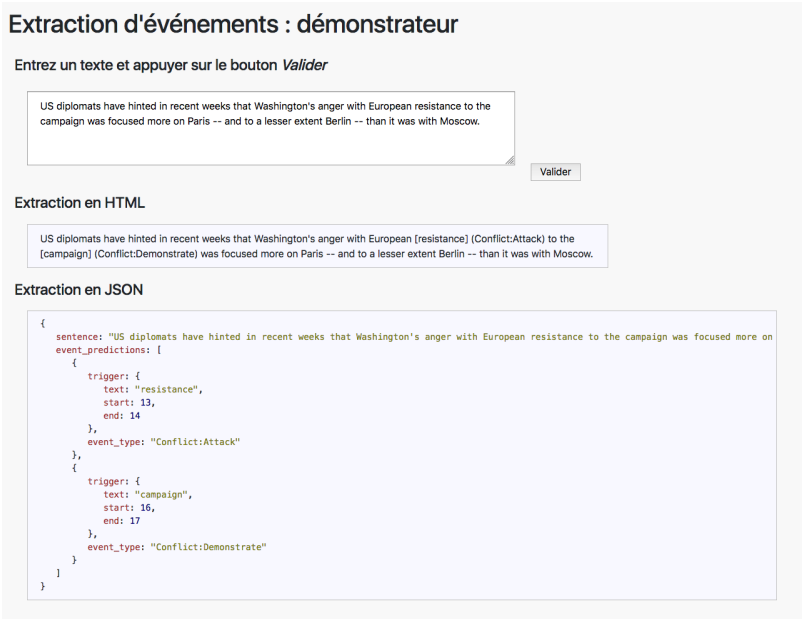
	Précision	Rappel	F-Mesure
Identification du trigger	0.34	0.82	0.48
Classification du trigger	0.35	0.84	0.49

TABLE 3.7 – Résultats de la maquette - Deuxième itération

Les résultats sont à nouveau encourageants. Ces modifications influent positivement à la fois sur la précision et le rappel, et la F-Mesure augmente de près de dix points. Ci-après (Figure 3.4 & Figure 3.5) sont présentés deux exemples de différences d'extraction entre les deux itérations.



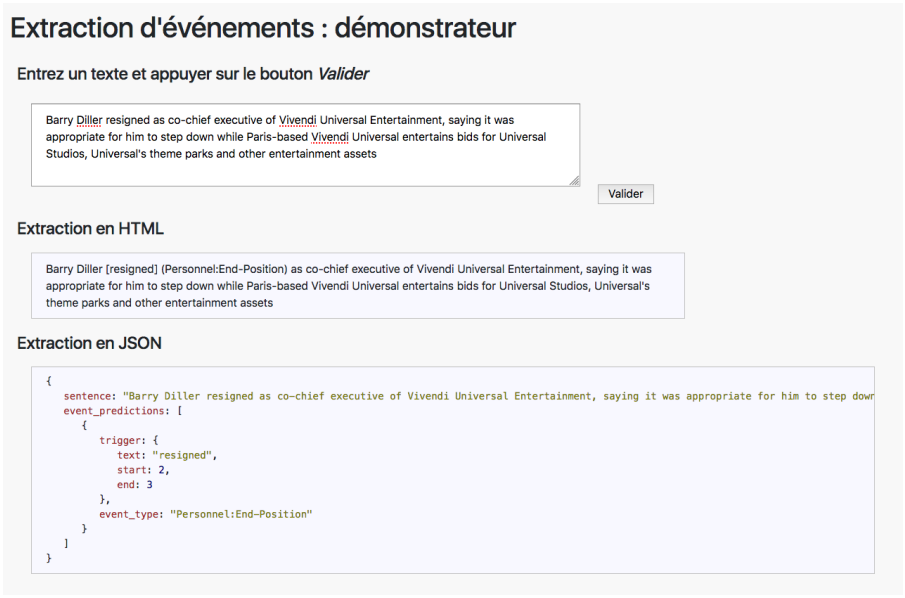
(a)



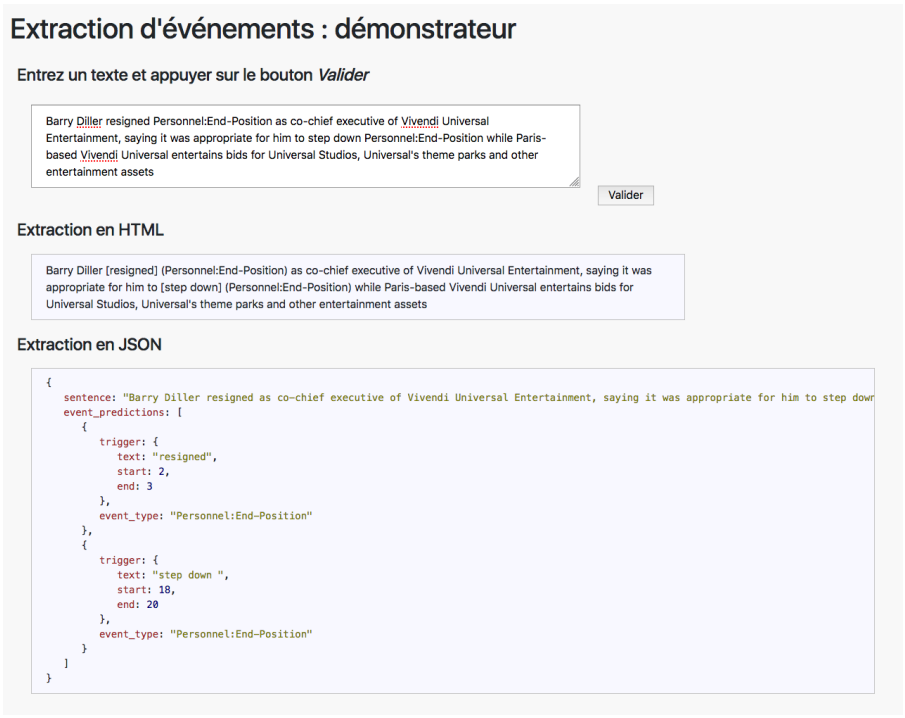
(b)

FIGURE 3.4 – TO HAVE - Comparaison d'extraction à la première (a) et deuxième (b) itération de la maquette

On voit ici (Figure 3.4) que l'auxiliaire *to have* était considéré comme un trigger dans la première itération. Après les modifications apportées lors de la deuxième itération, il n'est plus identifié comme un trigger, et ce, à raison.



(a)



(b)

FIGURE 3.5 – VERBES À PARTICULE - Comparaison d'extraction à la première (a) et deuxième (b) itération de la maquette

Dans cet exemple (Figure 3.5), le verbe à particule *to step down* n'était pas identifié comme un trigger lors de la première itération. Après les modifications apportées lors de la deuxième itération, il est désormais correctement identifié comme un trigger.

Ajouts des arguments et des rôles

L'ajout des motifs s'est fait à la troisième itération. Jusqu'ici, notre maquette vérifiait si dans une phrase p du texte t un couple lemme+étiquette morpho-syntaxique existait en tant que clé dans notre dictionnaire. Si tel était le cas, la valeur correspondant à cette clé, c'est-à-dire l'ensemble des types d'évènements possibles, était parcourue et le type d'évènement le plus fréquent était assigné au trigger, créant un formulaire. Lors de la troisième itération, nous avons enrichi le dictionnaire en assignant à chaque type d'évènement existant pour une clé l'ensemble des motifs d'entités possibles. Notons que le cas où l'évènement n'a pas d'arguments est également retenu dans notre maquette. Dans ce cas, c'est le motif ***NO_ARGS*** qui sera assigné.
Nous avons donc récupéré l'ensemble des entités présentes en tant qu'arguments dans chaque formulaire du corpus d'entraînement et de développement ; nous avons concaténé ces entités ainsi que le trigger pour former des motifs lisibles ; ces motifs respectent l'aspect séquentiel de la phrase puisqu'est indiqué l'indice de chaque entité au niveau phrastique. La phrase
Prime Minister Abdullah Gul resigned earlier Tuesday
(Le Premier ministre Abdullah Gul a démissionné mardi dernier)
engendre le formulaire suivant (Figure 3.6) :

```
{"golden-event-mentions": [
  {
    "trigger": {
      "text": "resigned",
      "start": 4,
      "end": 5
    },
    "arguments": [
      {
        "role": "Person",
        "entity-type": "PER",
        "text": "Prime Minister Abdullah Gul",
        "start": 0,
        "end": 4
      },
      {
        "role": "Time-Within",
        "entity-type": "TIM:time",
        "text": "Tuesday",
        "start": 6,
        "end": 7
      }
    ],
    "event_type": "Personnel:End-Position"}
```


FIGURE 3.6 – Formulaire d'évènement issu du corpus ACE 2005 au format JSON

Ce formulaire est intégré comme suit à notre dictionnaire :

— le trigger ***resign_VBD*** engendre

— l'évènement ***Personnel :End-Position***

— caractérisé par le motif d'entités ***PER_T_TIM :Time***

— & le motif de rôles ***PERSON_T_TIME-Within***.

Les valeurs de notre dictionnaire sont désormais enrichies des motifs d'entités et de rôles (Figure 3.7), classés eux aussi par pourcentage de fréquence. Les motif d'entités sont obtenus en concaténant les arguments (représentés par le type d'entités nommées correspondant) et le trigger (représenté par la lettre T). La concaténation, qui respecte l'ordre d'apparition des éléments dans la phrase, se fait à l'aide du caractère spécial underscore (_). Les motifs de rôles sont obtenus de manière similaire aux motifs d'entités, à la différence près que le type de rôle remplace le type d'entités nommées dans le motif construit.

```
war_NN {
    'event-type': ('Conflict:Attack', 100),
        'arguments': {
            'pattern': {
            ('GPE:Nation_T', 33.33): 'Attacker_T',
            ('T_GPE:Nation', 33.33): 'T_Place',
            ('PER_GPE:Nation_T', 33.33): 'Attacker_Place_T'
            }
        }
    }
```

FIGURE 3.7 – Extrait du dictionnaire à la troisième itération de la maquette

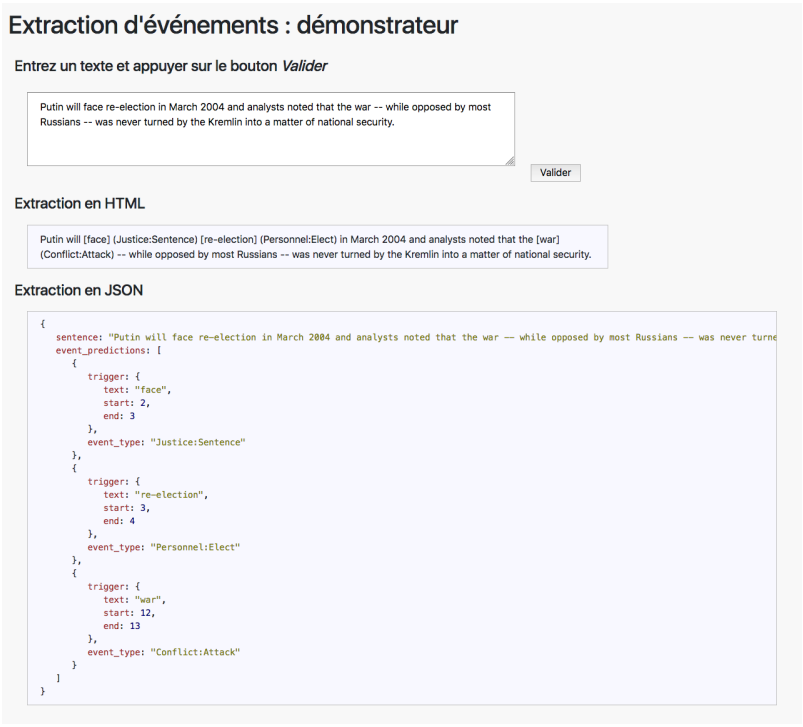
Afin d'activer la reconnaissance de ces motifs dans le texte d'entrée, nous avons adapté notre script d'extraction d'évènements pour transformer chaque phrase du texte en suite d'entités concaténées. Partant, nous procédons à la reconnaissance de motifs par une condition d'existence. Si le motif le plus fréquent est ***NO_ARGS***, le formulaire est automatiquement créé.

Les résultats de cette troisième itération sont présentés ci-dessous (Table 3.8).

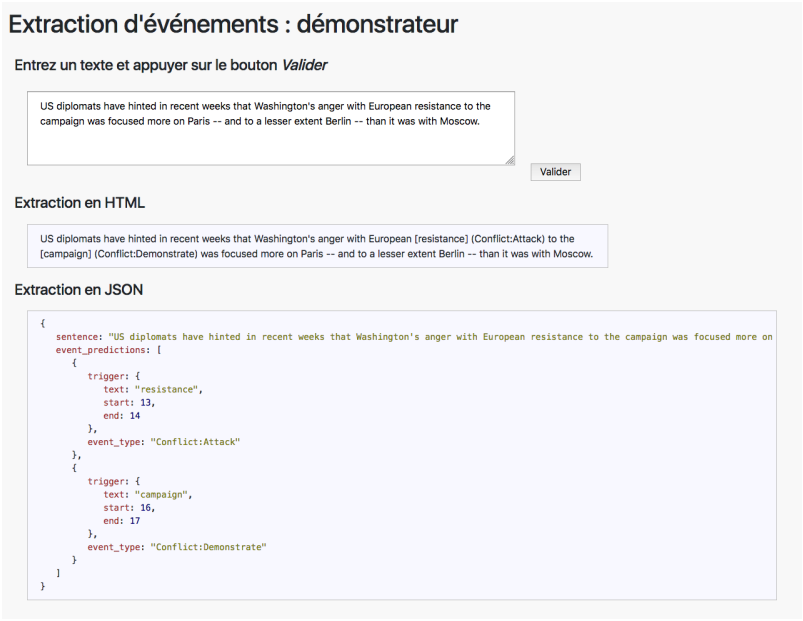
	Précision	Rappel	F-Mesure
Identification du trigger	0.45	0.65	0.53
Classification du trigger	0.44	0.63	0.52

TABLE 3.8 – Résultats de la maquette - Troisième itération

La baisse du rappel est compensée par l'augmentation de plus de dix points de la précision. L'ajout des motifs contribue donc à une détection plus précise des évènements.
Ci-après (Figure 3.8) est présenté un exemple de différence entre la deuxième et troisième itération.



(a)



(b)

FIGURE 3.8 – AJOUT DES MOTIFS - Comparaison entre la deuxième (a) et troisième (b) itération de la maquette

Dans cet exemple, nous voyons que l'ajout des motifs réduit d'un tiers le nombre d'évènements détectés. L'ajout des motifs contribue à une détection plus fine des évènements.

2.4 Améliorations possibles

Cette maquette d'extraction d'évènements constitue un premier pas vers la construction d'un outil d'extraction d'évènements à base de reconnaissance de motifs. Elle est bien sûr perfectible, et devra être enrichie. Ceci étant, l'évolution positive des résultats est encourageante et l'approche en reconnaissance de motifs de la maquette présente une marge de progression considérable. Cette approche présente également des possibilités de mise en oeuvre dans d'autres langues pour lesquelles existent des corpus annotés, par exemple l'arabe et le mandarin dans le cas du corpus ACE 2005.

Evaluation des arguments et des rôles

Notre script détecte les arguments de chacun des évènements et leur associe un rôle spécifique en accord avec le standard de notation ACE 2005. Une évaluation de l'extraction de ces arguments est en cours de réalisation.

Améliorations à l'aide des plongements lexicaux

L'utilisation des plongements lexicaux pourrait être de double utilité dans notre méthode. Premièrement, les plongements lexicaux pourraient être utilisés pour procéder à une classification sémantique de chaque phrase et/ou de chaque texte. Si certains triggers sont sémantiquement riches et peu ou pas ambivalents quant à l'évènement déclenché (*to kill*, *to injure*, *to resign*) d'autres restent ambigus si aucune information sur le contexte n'est disponible. Deuxièmement, les plongements lexicaux pourraient être utilisés pour créer, pour chacun des triggers extraits du corpus, une liste de triggers similaires au niveau morpho-syntaxique et sémantique. Notre maquette repose, pour l'instant, uniquement sur les triggers extraits des corpus d'entraînement et de développement du corpus ACE 2005. Les plongements lexicaux viendraient enrichir les résultats de cette extraction.

Mise en place d'une méthode similaire pour d'autres langues

Les ressources disponibles pour l'extraction d'évènements dans d'autres langues que l'anglais et le mandarin sont rares. De plus, ces ressources ne sont pas suffisantes pour les méthodes neuronales car les jeux de données sont déséquilibrés (Boros, 2018). Notre maquette ne repose pas sur une approche neuronale et sa méthodologie n'est pas propre à une langue donnée. Par conséquent, elle peut être transposable à d'autres langues par la création de corpus annotés de manière semi-automatique et d'extraction de motifs à partir de ces corpus.

Chapitre 4

Conclusion & perspectives

L'extraction d'évènements est une tâche capitale dans le processus de veille car elle permet de structurer l'information recueillie de manière extrêmement précise et d'en faciliter la manipulation et l'utilisation. L'objet du stage au sein de l'équipe R&D de Bertin IT était triple : rédiger un état de l'art sur l'extraction d'évènements, parcourir les outils consacrés à l'extraction d'évènements disponibles en Open Source, et mettre en place une maquette d'extraction d'évènements.

La rédaction de l'état de l'art destiné à l'entreprise a largement contribué à l'élaboration de ce rapport, et sa structure a guidé le travail réalisé tout au long du stage. La définition de l'extraction d'évènements, le relevé de différentes approches adoptées dans le domaine et l'énumération des nombreuses campagnes d'évaluation en extraction d'évènements ont donné à l'entreprise une première vision de la tâche et nous ont permis de tester des outils disponibles en Open Source et d'élaborer une maquette en extraction d'évènements.

Les outils récents d'extraction d'évènements disponibles en Open Source sont majoritairement issus de méthodes faisant appel à des réseaux de neurones. Le test de certains de ces outils nous a amenées à penser que ces méthodes présentent de nombreux désavantages pour une utilisation industrielle. Elles offrent peu de contrôle sur leur fonctionnement et leur résultat, les temps d'apprentissage sont longs et la construction de modèles pour obtenir des résultats optimaux est difficile.

Partant, nous avons proposé une maquette d'extraction d'évènements reposant sur une approche de reconnaissances de motifs d'entités linguistiques issus du corpus ACE 2005. Grâce à l'adoption d'un processus itératif, nous avons bénéficié d'un repère en terme de score tout au long de l'élaboration de la maquette. Celle-ci a été enrichie au fur et à mesure, et soumise à chaque itération à notre script d'évaluation.

Les perspectives pour l'évolution de la maquette sont positives. Notre approche inspirée des méthodes en reconnaissance de motifs suppose une marge de progression considérable. En outre, les résultats de notre script d'évaluation ont montré une progression significative des résultats. Cette progression possible et l'évolution positive des résultats sont encourageants et nous invitent à améliorer et enrichir la maquette, et à la développer dans d'autres langues.

Liste des figures

Liste des tableaux

Bibliographie

Ahn, D. (2006). The stages of event extraction, In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, Sydney, Australia, Association for Computational Linguistics. https://www.aclweb.org/anthology/W06-0901

Arnulphy, B. (2012). *Désignations nominales des événements : étude et extraction automatique dans les textes* (thèse de doct.). Université Paris Sud - Paris XI, 2. https://tel.archives-ouvertes.fr/tel-00758062

Bengio, Y. (2019). *La révolution de l'apprentissage profond*. Récupérée 6 juillet 2020, à partir de https://interstices.info/la-revolution-de-lapprentissage-profond/

Besançon, R. (2017). Problématiques et approches pour la détection d'événements dans des textes. *Séminaire IRIT*, 1-61. https://www.irit.fr/IRIS-site/images/seminairs/Besancon2017.pdf

Björne, J. & Salakoski, T. (2018). Biomedical Event Extraction Using Convolutional Neural Networks and Dependency Parsing, In *Proceedings of the BioNLP 2018 workshop*, Melbourne, Australia, Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2311

Boroş, E. (2018). *Neural Methods for Event Extraction* (thèse de doct.). Université Paris-Saclay. https://tel.archives-ouvertes.fr/tel-01943841

Calvo Rigual, C., Minervini, L. & Thibault, A. (2013). Les verbes à particule d'origine anglaise en français louisianais. *Actes du XVIIè Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*. http://www2.atilf.fr/cilpr2013/actes/section-11/CILPR-2013-11-Rottet.pdf

Cao, K., Li, X., Fan, M. & Grishman, R. (2015). Improving Event Detection with Active Learning, In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, INCOMA Ltd. Shoumen, BULGARIA. https://www.aclweb.org/anthology/R15-1010

Carreras, X. & Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task : Semantic Role Labeling, In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan, Association for Computational Linguistics. https://www.aclweb.org/anthology/W05-0620

Chen, Y., Liu, S., Zhang, X., Liu, K. & Zhao, J. (2017). Automatically Labeled Data Generation for Large Scale Event Extraction, In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Vancouver, Canada, Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1038

CORDIS. (2017). *Building structured event indexes of large volumes of financial and economic data for decision making*. Récupérée 22 juillet 2020, à partir de https://cordis.europa.eu/project/id/316404

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. & Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation, In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf

Ello. (2009). *Frame Semantics*. Récupérée 9 juillet 2020, à partir de http://www.ello.uos.de/field.php/

eMarketing. (2020). *Veille stratégique (ou veille concurrentielle)*. Récupérée 29 août 2020, à partir de https://www.e-marketing.fr/Definitions-Glossaire/Veille-strategique-veille-concurrentielle-238998.htm

Framenet. (2020). *What is FrameNet ?* Récupérée 2 juillet 2020, à partir de https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet

Grishman, R. (2010). The Impact of Task and Corpus on Event Extraction Systems. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. http://www.lrec-conf.org/proceedings/lrec2010/pdf/565_Paper.pdf

Grishman, R. & Sundheim, B. (1995). Design of the MUC-6 Evaluation. *Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. https://doi.org/10.3115/1072399.1072401

Grishman, R. & Sundheim, B. (1996). Message Understanding Conference- 6 : A Brief History, In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C96-1079

Haris, D. (2014). *DARPA is working on its own deep-learning project for natural-language processing*. Récupérée 3 juillet 2020, à partir de https://gigaom.com/2014/05/02/darpa-is-working-on-its-own-deep-learning-project-for-natural-language-processing/

Hayes, W. (2018). What is BEL ? Récupérée 22 juillet 2020, à partir de https://medium.com/biodati/what-is-bel-8df1a549760f

Horev, R. (2018). BERT Explained : State of the art language model for NLP. Récupérée 30 juillet 2020, à partir de https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Hsi, A., Yang, Y., Carbonell, J. & Xu, R. (2016). Leveraging Multilingual Training for Limited Resource Event Extraction, In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, Osaka, Japan, The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/C16-1114

Jean-Louis, L. (2011). *Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances* (thèse de doct.). Université Paris Sud - Paris XI. https://tel.archives-ouvertes.fr/tel-00686811

Kodelja, D., Besancon, R. & Ferret, O. (2017). Représentations et modèles en extraction d'événements supervisée. *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017)*. https://hal.archives-ouvertes.fr/hal-01561986

Ldc. (2005). *ACE (Automatic Content Extraction) English Annotation Guidelines for Events - Version 5.4.3 2005.07.01*. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

Li, Q., Ji, H. & Huang, L. (2013). Joint Event Extraction via Structured Prediction with Global Features, In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Sofia, Bulgaria, Association for Computational Linguistics. https://www.aclweb.org/anthology/P13-1008

Liu, X., Huang, H. & Zhang, Y. (2019). Open Domain Event Extraction Using Neural Latent Variable Models, In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1276

Liu, X., Luo, Z. & Huang, H. (2018). Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1156

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit, In *Association for Computational Linguistics (ACL) System Demonstrations*. http://www.aclweb.org/anthology/P/P14/P14-5010

NetOwl. (2019). *What is Event Extraction ?* Récupérée 2 juillet 2020, à partir de https://www.netowl.com/what-is-event-extraction/

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe : Global Vectors for Word Representation, In *Empirical Methods in Natural Language Processing (EMNLP)*. http://www.aclweb.org/anthology/D14-1162

Polcon. (2018). *Shared Task on Protest Event Mining*. Récupérée 9 juillet 2020, à partir de https://www.eui.eu/Projects/POLCON/Shared-Task-on-Protest-Event-Mining

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages, In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. *AAAI'96 : Proceedings of the thirteenth national conference on Artificial intelligence*, *2*, 1044-1049. https://doi.org/10.5555/1864519.1864542

Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S. & Grilheres, B. (2013). Extraction et agrégation automatique d'événements pour la veille en sources ouvertes : du texte à la connaissance, In *IC - 24èmes Journées francophones d'Ingénierie des Connaissances*, Lille, France. https://hal.archives-ouvertes.fr/hal-01024341

Tac. (2017). *TAC KBP 2016 Event Track*. Récupérée 9 juillet 2020, à partir de https://tac.nist.gov/2016/KBP/Event/index.html

Webster, J. J. & Kit, C. (1992). Tokenization As The Initial Phase In NLP, In *COLING*.

Wikipedia. (2003). *Information Extraction*. Récupérée 3 juillet 2020, à partir de https://en.wikipedia.org/wiki/Information_extraction

Wikipedia. (2006). *Message Understanding Conference*. Récupérée 8 juillet 2020, à partir de https://en.wikipedia.org/wiki/Message_Understanding_Conference

Xiang, W. & Wang, B. (2019). A Survey of Event Extraction From Text. *IEEE Access*, *7*, 173111-173137. https://doi.org/10.1109/ACCESS.2019.2956831