Induction non-supervisée de schémas d'évènements à partir de textes journalistiques

Swen Ribeiro

université PARIS-SACLAY

Induction non-supervisée de schémas d'événements à partir de textes journalistiques

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 : sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Unité de recherche : Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 10 Mars 2020, par

Swen RIBEIRO

Composition du Jury

Karine ZEITOUNI Professeur, Université de Versailles St Quentin en Yvelines	Présidente
Antoine DOUCET Professeur, Université de la Rochelle	Rapporteur & examinateur
Philippe LANGLAIS Professeur, Université de Montréal	Rapporteur & examinateur
Kata GÀBOR Maître de conférences, INALCO	Examinatrice
Xavier TANNIER Professeur, Sorbonne Université, LIMICS	Directeur de thèse
Olivier FERRET Ingénieur Chercheur, CEA-LIST	Co-Encadrant & examinateur

Thèse de doctorat

NNT : 2020UPASS059

Résumé

L'événement est un concept central dans plusieurs tâches du Traitement Automatique des Langues, en dépit de l'absence d'une définition unifiée de ce que recouvre cette notion. Le traitement des événements s'est structuré sous l'égide des campagnes d'évaluation MUC (*Message Understanding Conference*), qui fournissaient des structures de référence appelées schémas (*templates* en anglais), se présentant sous la forme d'un titre et d'une collection d'arguments (*slots*), chacun représentant un élément caractéristique de l'événement décrit (par exemple l'*épicentre* d'un séisme). La création de ces schémas requiert une connaissance experte et est donc longue, coûteuse et difficile à étendre à un large ensemble de domaines de spécialité.
En parallèle de ces travaux, la quantité de données produites par les individus et les organisations a crû de manière exponentielle, ouvrant des perspectives applicatives inédites. Cette croissance a notamment favorisé l'essor d'un nouveau paradigme journalistique appelé journalisme de données (*data-journalism* en anglais).
Le présent travail se propose d'induire, à partir d'un grand volume de texte journalistique et sans supervision, des représentations synthétiques d'événements journalistiques comparables aux *templates* des campagnes MUC, dans l'objectif de faciliter l'exploitation de grandes masses de données par des journalistes des données. Pour ce faire, nous suivons une approche ascendante divisée en trois grandes étapes. Dans la première étape, nous groupons ensemble les nombreuses mentions textuelles relatant la même réalisation d'un événement, identifiée dans le temps et l'espace et appelée instance. La deuxième étape vise à s'abstraire des caractéristiques spatio-temporelles de chaque instance pour les grouper en grands types d'événements. Enfin, la dernière étape de cette contribution vise à extraire les éléments caractéristiques de chaque type d'événement induit afin d'en proposer une représentation synthétique assimilable à un schéma d'événement.

Abstract

Events are central in many Natural Language Processing tasks, despite the lack of a unified definition for the concept. The field of event processing took off with the MUC evaluation campaigns that provided participants with reference structures called templates. These templates were composed of a title (the name of the event) and several slots, i.e specific and atomic pieces of data about the event. Creating these templates is an expert task and therefore costly, painstaking and hard to extend to new domains.

Meanwhile, the amount of data produced by individuals and organizations has grown exponentially, opening unprecedented perspectives of applications. In the journalistic domain, it fueled the development of a new paradigm called data-journalism.

In this work, we aim at inducing synthetic representations of events from large textual journalistic corpora. These representations would be comparable to MUC templates and used by data-journalists to explore large textual news datasets. To this end, we propose a bottom-up approach composed of three main steps. The first step clusters several textual mentions of a same particular event (i.e. tied to a time and place) to identify distinct *instances*. The second step groups these instances together based on more abstract features to infer event types. Finally, the third and last step extracts the most salient elements of each type to produce the synthetic, template-like structure we are looking for.

Remerciements

La thèse est souvent perçue (et décrite) comme l'aboutissement d'un travail individuel. Certaines diront simplement que “c'est pas faux". D'autres diront que certes le travail est solitaire mais la réflexion qui y mène l'est plus rarement et appuieront cette affirmation en citant les nombreuses thèses s'ouvrant sur une page de remerciements (voire plusieurs pour les plus épris d'exhaustivité). Je ne dérogerai pas à la règle et m'efforcerai ici de faire faire honneur à toutes celles et tout ceux qui ont fait de ma thèse une période si marquante de ma vie.

Je souhaite en premier lieu remercier Antoine Doucet et Philippe Langlais pour avoir accepté le rôle de rapporteurs de ce travail, ainsi que Kata Gábor et Karine Zeitouni pour avoir endossé celui d'examinatrices. Leurs remarques et leurs questions m'ont permis de mesurer le travail accompli, une mise en perspective très importante pour moi et pour laquelle ils ont toute ma gratitude.

Je souhaite également remercier Xavier Tannier et Olivier Ferret, qui m'ont guidé pendant ces quatre années. Ils m'ont accompagné jusqu'au bout de cette expérience avec autant de patience que de constance, veillant à ce que je puisse toujours me consacrer à mon travail dans les meilleures conditions. Je souhaite également remercier Thomas Lavergne, sans qui je n'aurais jamais vécu cette expérience aujourd'hui si importante pour moi.

Au risque d'en surprendre certains, la thèse, et notamment ses difficultés, ne se résument pas à la recherche rigoureuse de résultats repoussant les limites de la connaissance humaine. De nombreuses considérations bien plus terre-à-terre peuvent s'imposer au doctorant. Je souhaite donc remercier toute l'équipe de soutien à la recherche du LIMSI, des gestionnaires à l'équipe informatique, qui veille tous les jours à rendre ces charges les plus légères possible et contribue à faire du laboratoire un lieu dans lequel on se rend chaque jour avec plaisir.

La formation à la recherche ne se fait pas dans une salle de classe. Chaque échange est un enseignement et si Xavier et Olivier ont indéniablement contribué activement à ce processus, il serait injuste d'oublier tous les collègues du LIMSI qui ont prolongé cet apprentissage en dehors du cadre de ma thèse. Je remercie en particulier Patrick

Paroubek et Michael Filhol pour la richesse de nos conversations et tout ce qu'elles m'ont apporté, sur le monde et sur la science.

Sanjay Kamath, Zheng Zhang, Arthur Boyer, Yuming Zhai, Arnaud Ferré, Julien Tourille et Marine Delaborde ont quotidiennement éclairé mes jours de thèse et pour cela je leur exprime toute ma gratitude. Ils ont toutefois fait le choix regrettable de finir leurs thèses avant moi ou de quitter prématurément le laboratoire pour d'autres horizons, un faux pas heureusement compensé par leur capacité à me supporter durant le temps que nous avons partagé (en espérant qu'il n'y ait pas de cause à effet). À l'inverse, Hicham El Boukkouri est arrivé trop tard à mon goût mais à point nommé pour participer à l'embellissement du présent travail. Un immense merci à tous pour ces moments passés ensemble. Un grand merci également à Charlotte Rudnik, sans qui les derniers chapitres de cette thèse seraient beaucoup plus courts et à Matthieu Labeau et Franck Burlot, premiers collègues alors que je n'étais qu'un stagiaire insouciant.

Si mes directeurs ont veillé à maintenir un contexte de travail le plus favorable possible, mes parents se sont attachés à me faire éviter de nombreux obstacles de tout le reste de la vie quotidienne. Leur soutien infaillible m'a permis de me consacrer pleinement à mon travail de thèse et toute cette énergie ne fut pas de trop pour arriver au bout de ce voyage. Si la thèse a fait de moi un chercheur, ils ont fait de moi la personne que je suis, celle qui a réussi cette thèse, et mon amour pour eux ne saurait se résumer en quelques lignes.

Enfin, les mots ne suffiront pas à exprimer l'amour et la reconnaissance que j'éprouve à l'égard de Lucie Gianola. Elle fut certainement la plus grande découverte de ma thèse et sa présence à mes côtés fut la source des plus précieux enseignements que j'ai tirés de cette période. Mon plus grand bonheur est de savoir que nous arpentons ensemble une route encore longue. Je n'aurais jamais fini de te remercier.

Table des matières

1 Introduction

> *Oui, eh ben si vous vouliez du captivant, fallait peut-être me faire lire autre chose[. . .]*
>
> — **Jean-Robert Lombard**
> Kaamelott, Livre IV, Le Vice De Forme

1.1 Contexte général

Les travaux présentés dans ce manuscrit visent à appliquer des techniques de Traitement Automatique des Langues (TAL) à des problématiques de journalisme de données (*data-journalism* en anglais), un mouvement qui redéfinit la production de contenu journalistique tout en conservant le but initial de la profession, à savoir informer le public de l'état du monde.
Le journaliste "traditionnel" accomplit cette tâche en collectant, auprès de sources fiables, les faits qui constitueront la teneur de cet éclairage, en vérifiant ces faits et en les structurant dans un récit. Le journaliste de données collecte quant à lui des données, les traitent et les analysent à l'aide d'outils issus de différents champs scientifiques (statistiques, sciences humaines. . .). Ce sont les résultats de cette analyse qui constituent l'information portée à la connaissance du public.
Si certains auteurs font remonter le journalisme de données jusqu'au début du XIX^{e} siècle, la pratique n'émerge vraiment qu'aux États-Unis dans les années 50, sous le nom de *Computer Assisted Reporting*, puis *precision journalism*. La pratique se structure dans les années 70 et 80 en opposition au "nouveau journalisme" (*new journalism* en anglais), qui donnait une grande place à la forme au détriment du fond, en développant le recours à des ressorts narratifs issus de la fiction. Elle est aujourd'hui une pratique reconnue et bien implantée dans le monde anglo-saxon et se développe dans le reste du monde [1].

Le journalisme de données met en exergue le couplage que le journalisme "traditionnel" effectue entre sources et données : si la source tient à être protégée, la déontologie exige que son anonymat soit préservé, ce qui peut rendre difficile le

1. Informations extraites à partir de : `https://datajournalism.com/read/handbook/one/introduction/data-journalism-in-perspective`.

partage des informations entre journalistes ou la vérification des faits et discréditer la source ou le travail du journaliste (rendant par la même la recherche du *scoop* plus laborieuse, puisqu'elle implique de prendre contact avec des sources bien placées, qui tiendront forcément à leur anonymat et fourniront des informations difficiles à vérifier puisque difficiles à obtenir en premier lieu). La valeur ajoutée du journaliste de données se situe quant à elle plutôt au niveau de l'analyse et du sens qui est fait des données, ce qui permet le partage de cette dernière (et la vérification des analyses proposées) et facilite la protection des sources. De fait, le journalisme de données accorde une grande place à la transparence, s'inscrivant pleinement dans la mouvance sociétale de l'*Open Data*, c'est-à-dire la mise à disposition du public par les institutions des données qu'elles exploitent et produisent. Cette mouvance générale a également permis à la pratique de dépasser le cadre du journalisme d'investigation qui l'a vu naître, puisqu'elle constitue un moyen de faire face aux défis de l'ère du Web et des réseaux sociaux, notamment les grandes quantités d'informations contradictoires qu'ils engendrent, le *fact-checking* étant un exemple de journalisme de données. Il est important de remarquer que le journalisme de données est une pratique journalistique qui doit être envisagée comme un complément et non un substitut au journalisme "traditionnel", ayant pour objectif d'adapter les processus de production de l'information à de nouveaux types de sources.

Une caractéristique définitoire de cette pratique est de travailler sur des ensembles substantiels de données, nécessitant le recours à des outils d'analyse pour permettre aux équipes de journalistes d'accomplir leur tâche dans les délais qui leur sont impartis. C'est dans cette composante "technophile" que s'inscrit le projet ANR ASRAEL, en partenariat avec l'Agence France Presse (AFP), dans le périmètre duquel les présents travaux ont été réalisés.
En tant qu'agence de presse, l'AFP dispose de grands volumes de données, alimentés par un réseau mondial de bureaux locaux, réactifs et fournissant une information fiable, servant de matière première à la production de contenu pour de nombreux autres organes de presse. Dans ce contexte, nos travaux ont visés à réaliser un système de représentation d'événements dans le but de fournir aux journalistes de l'AFP des moyens de tirer parti des masses d'informations se trouvant à leur disposition (qu'elles soient extérieures et ouvertes ou internes) pour traiter l'événement dans l'ensemble de ses dimensions et selon le paradigme du journalisme de données.
Notre objectif n'est pas d'assister le journaliste dans les tâches de collecte d'informations de première main, de rédaction de dépêches ou de création d'autres contenus portant sur des réalisations particulières d'événements dans le réel mais de permettre la production de contenus de plus large échelle, telles que des frises chronologiques ou des cartes interactives, domaine où l'agence de presse peut pleinement tirer parti de sa position d'acteur central du transit d'information.

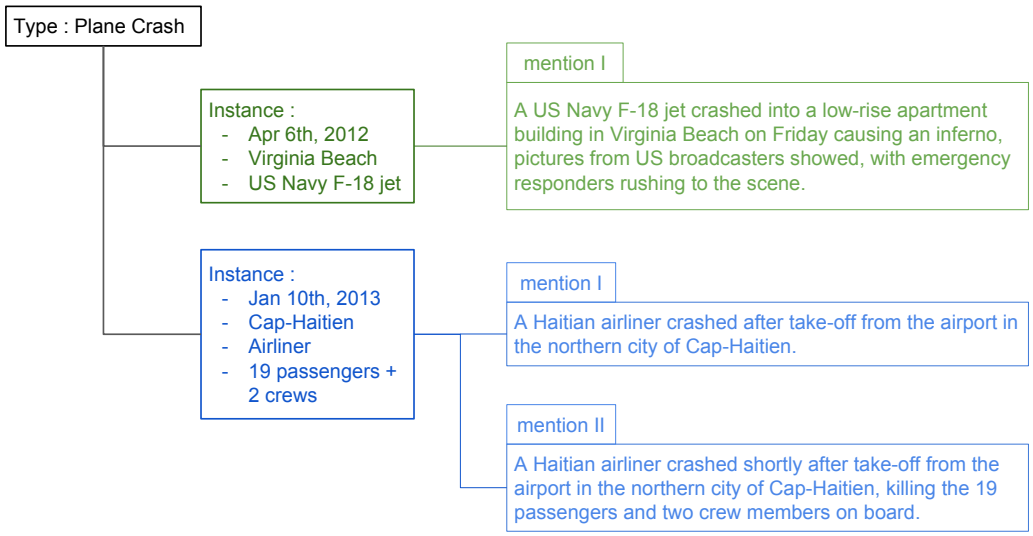


Fig. 1.1: Illustration de la hiérarchie structurant les notions de mentions, instances et types d'événements : plusieurs mentions peuvent décrire une même instance et différentes instances peuvent être regroupées sous un même type.

1.2 Problématique de la thèse

Comme nous l'avons exposé dans la section précédente, le journalisme de données exploite de nombreux types de données différents, particulièrement sous des formats numériques et structurées. Néanmoins, le travail présenté ici se focalisera sur le traitement de données textuelles.
Notre objectif est d'induire des représentations génériques d'événements à partir de textes décrivant leurs réalisations particulières, selon le processus visible en figure 1.1, afin d'obtenir des schémas d'événements, dont deux exemples sont illustrés en figure 1.2 [2].

Le partenariat avec l'AFP dans le contexte du projet ASRAEL nous a donné accès à un grand volume de leurs dépêches. Comme nous le verrons dans la suite de ce manuscrit, ces données se distinguent d'autres contenus journalistiques par un certain nombre de caractéristiques que nous essayons d'exploiter. Cette volumétrie constitue néanmoins en elle-même un élément déterminant dans la structuration de notre approche, puisqu'elle pose à la fois des contraintes techniques dont nous devons tenir compte afin de produire un système exploitable en pratique tout en constituant un avantage dont nous souhaitons tirer profit.

L'induction de schémas d'événements implique l'identification précise d'éléments caractéristiques de chaque type d'événement décrit dans cette masse de données. Cet

2. Nous définissons le concept de schéma d'événement ainsi que d'autres concepts clés de nos travaux en section 1.3.

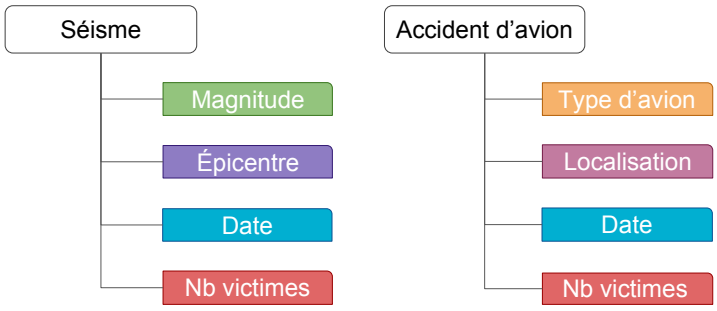


Fig. 1.2: Deux schémas possibles représentant les événements “Séisme" à gauche et “Accident d'avion" à droite. On remarquera que plusieurs arguments différencient ces deux événements (Magnitude et Type d'avion) mais également qu'ils en partagent, comme Nombre de victimes (ou Date, dont on peut s'attendre à ce qu'ils soient présents pour tous les schémas, conformément à notre définition d'un événement). On peut remarquer enfin les arguments Épicentre et Localisation qui, sous une dénomination différente (propre au type de l'événement), décrivent une caractéristique comparable.

objectif de produire une synthèse structurée à partir d'un vaste ensemble de données non-structurées situe ce travail à l'interface entre les domaines de la Recherche d'Information (RI) et de l'Extraction d'Information (EI).
La Recherche d'Information (RI) est un champ de recherche visant à répondre à un besoin informationnel, exprimé par un utilisateur, par l'exploitation d'un corpus de données. Fondamentalement, un système de RI est associé à un corpus que l'utilisateur ne peut appréhender par ses propres moyens. Cette impossibilité peut venir de différents facteurs (type des données, technologie des supports d'information, besoin de fusionner des informations de types différents, . . .) mais la problématique principale qui a fait émerger ce champ et le structure depuis ses débuts est celle du volume : les systèmes de RI sont le plus souvent conçus pour organiser et permettre l'accès à un volume de données trop important pour être traité frontalement par ses utilisateurs, la satisfaction de chacun de leur besoin d'information revenant pour ces derniers à “chercher une aiguille dans une botte de foin". Pour ce faire, la RI distingue deux phases : la première, nommée “indexation", est un préalable à toute utilisation du système et consiste à organiser le contenu du corpus pour permettre un accès rapide et ciblé aux informations qu'il contient. La seconde, appelée “interrogation", correspond à la résolution du besoin informationnel de l'utilisateur. Elle commence par l'expression de ce besoin au travers d'une requête interprétable pour le système de RI et s'appuie sur le produit de l'indexation pour rechercher les éléments du corpus susceptibles de la satisfaire. D'autres traitements peuvent suivre avant de présenter ces résultats à l'utilisateur, comme leur organisation par ordre de pertinence.

L'Extraction d'Information (EI) se focalise sur la structuration d'informations à partir de sources non-structurées ou semi-structurées. Elle complète et étend la RI

en cela qu'elle suppose l'existence d'une structure latente dans les données dont l'exploitation permettrait de synthétiser les informations qu'elles contiennent. Dans ce cadre, un système de RI sélectionne un sous-ensemble du corpus de travail qui répond au mieux au besoin informationnel de l'utilisateur et un système d'EI identifie et structure les informations les plus remarquables de ce sous-ensemble.

Dans ce travail, nous abordons l'induction de schémas d'événements d'abord sous l'angle de la RI, en organisant notre corpus initial en blocs d'événements via deux phases pouvant être apparentée à de l'indexation, puis en allant chercher dans ces blocs leurs éléments caractéristiques individuels afin d'en produire une synthèse structurée, ce qui rejoint les objectifs de l'EI. Réaliser ce travail d'induction avec le moins de supervision possible, et idéalement aucune, constitue une problématique centrale de notre travail, à la résolution difficile mais nécessaire en raison tout d'abord de la difficulté intrinsèque pour les humains à définir de telles structures *a priori*, qui se trouve accentuée par la grande variété de formes (tant en termes de conditions de réalisation que de description textuelle) que peut prendre chaque instance d'événement à travers le temps et les sources. De plus, l'AFP étant une agence de presse généraliste, cette volumétrie de texte décrit une variété d'événement tout aussi large dont la description formelle par des humains, même de façon superficielle, impliquerait des coûts humain, temporel et financier inacceptables.

Dans la perspective du journalisme de données, l'induction de schémas d'événements revêt néanmoins un intérêt fort car ces structures ont le potentiel d'inscrire n'importe quelle production journalistique (textuelle dans notre cas) dans un ensemble plus global (le schéma) aux caractéristiques explicites pour l'humain permettant de traiter le ou les événements sous-jacents selon différents angles.

Dans ce travail, nous proposons une solution s'appuyant successivement sur trois niveaux de granularité (introduits en section 1.3) pour qualifier l'événement, à savoir la mention, l'instance et le type. À l'instar de travaux précédents, nous suivons une approche en *pipeline* dans laquelle chaque module passe d'un niveau à l'autre. Nous commencerons, dans le chapitre 2, par dresser un panorama historique des travaux antérieurs aux nôtres et ayant pris pour sujet d'étude l'événement. Les chapitres suivants décrivent notre contribution : le chapitre 3 détaille comment nous exploitons les mentions d'événements issues de corpus hétérogènes pour identifier des instances d'événements. Le chapitre 4 s'attarde quant à lui sur la structuration de ces instances en types d'événements. En dernier lieu, le chapitre 5 présente la solution retenue pour extraire les éléments constitutifs de ces structures, reposant sur des regroupements de documents. Nous terminerons par exposer des perspectives d'amélioration ou d'application possible de cette solution.

1.3 Définitions

Comme le soulignent FILATOVA et al. (2006), la terminologie en matière de traitement de l'événement peut parfois être ambiguë en raison des différents paradigmes qui se sont mis en place indépendamment les uns des autres. C'est pourquoi il nous apparaît très important de poser dès maintenant quelques définitions.

Événement : nous considérons un événement comme "quelque chose se produisant quelque part", c'est-à-dire un changement d'état du monde dans un contexte spatial et temporel délimité. Nous ajouterons que ce changement doit être suffisamment significatif (par sa nature ou son intensité) pour être relayé par des canaux journalistiques ;

Mention (d'événement) : nous désignerons par mention une description textuelle d'un événement, indépendamment de la longueur de celle-ci. Chaque événement peut faire l'objet de plusieurs mentions mais une mention ne concerne qu'un et un seul événement ;

Type (d'événement) : un type d'événement se caractérise par un terme ou une expression, souvent nominaux, indiquant la nature de l'événement considéré. "Séisme", "Attentat à la bombe" ou "Rencontre sportive" sont des exemples de types d'événements. La granularité de ces types (rencontre sportive *vs* match de tennis) n'est pas fixée et représente un des enjeux de notre travail ;

Instance (d'événement) : chaque réalisation dans le monde réel, c'est-à-dire dans un contexte spatio-temporel particulier, d'un type d'événement est considérée comme une instance. Par exemple, les séismes du 25 Avril 2015 au Népal et du 16 Avril 2015 au Chili sont deux instances distinctes d'un même type d'événement ;

Schéma (d'événement) : un schéma d'événement est une représentation générique et synthétique de l'événement. Elle se compose d'un noyau, qui indique le type de l'événement, et d'un nombre variable d'arguments, qui sont les éléments caractéristiques de cet événement. Deux exemples graphiques de schémas d'événements sont proposés en figure 1.2.

En complément de ces explications, une représentation de la hiérarchie mention > instance > type est visible en figure 1.1. La figure 1.3 illustre le passage de la mention au schéma, qui est l'objectif du projet. On y remarquera que certains éléments sont communs à chaque mention (dans l'exemple présenté, le type d'avion impliqué est indiqué dans tous les cas, ainsi que le lieu de l'accident), tandis que d'autres diffèrent (un bilan des victimes n'est mentionné que dans le texte en bleu). Tous sont néanmoins pertinents pour la caractérisation du type d'événement considéré. Nous souhaitons exploiter de grands volumes de mentions de ce genre pour en constituer

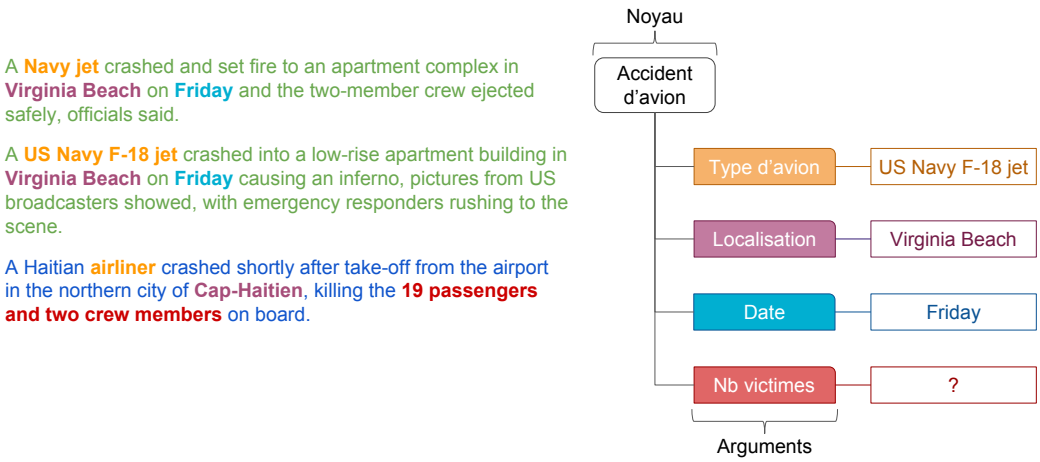


Fig. 1.3: Exemple de synthèse de mentions d'événements en schéma. Sur la gauche, les deux mentions en vert décrivent la même instance d'événement, celle en bleu relate une instance différente mais du même type, à savoir un accident d'avion. À droite, un exemple de schéma associé, avec son noyau, ses arguments et les mentions associées dans l'instance décrite en vert.

des ensembles dont chacun se focalisera sur un type d'événement particulier et dans lesquels chaque argument pourra être identifié car il sera représenté dans plusieurs mentions.

1.4 Présentation des ressources et hypothèses de travail

1.4.1 Remarques préliminaires

Nos travaux ont porté sur des données issues de corpus en langue anglaise. Pour cette raison, les figures de ce manuscrit s'appuieront sur des textes en anglais.

Par ailleurs, nous avons utilisé plusieurs fois dans nos travaux des méthodes de *clustering*. Bien que ce terme soit un anglicisme, nous l'utiliserons dans la suite de ce document car cet emploi nous a semblé plus judicieux que celui de ses équivalents français "regroupement" ou "agglomération", en raison de leur ambiguïté, notamment entre le processus (clustering) et le résultat (clusters).

1.4.2 Corpus AFP

Dans le cadre du projet ASRAEL, l'AFP a mis à notre disposition l'ensemble des dépêches émises sur les dernières années. Chaque dépêche se compose comme tout

contenu journalistique d'un titre, suivi d'un chapeau, puis du corps de la dépêche. Chacun de ces éléments est produit sous des contraintes rédactionnelles plus ou moins fortes.

D'abord, le titre doit résumer la nouvelle que relaie la dépêche, en quelques mots seulement. Puis le chapeau doit donner au lecteur tous les éléments nécessaires à la compréhension de la nouvelle. Enfin, le corps de la dépêche doit décrire en détail la nouvelle en y incluant des informations de contexte général. Ce dernier est souvent enrichi à mesure que des éléments de la nouvelle s'ajoutent ou se précisent, ce qui en fait la partie la plus "variable" de la dépêche pour la description d'une même nouvelle.

Le chapeau est probablement l'espace le plus contraint du fait de la nécessité de donner un maximum d'informations au lecteur en un minimum de mots. Cette partie constitue le premier paragraphe de la dépêche et consiste généralement en une seule phrase qui doit répondre aux "cinq W", faisant référence aux 5Ws anglo-saxons, *Who* (Qui), *What* (Quoi), *When* (Quand), *Where* (Où), *Why* (Pourquoi), c'est-à-dire les éléments définissant les circonstances d'un événement.

À l'inverse, le corps de la dépêche est un espace qui permet de détailler l'événement dans toute sa complexité, ce qui implique la description d'événements connexes et l'émergence d'un réseau d'événements plus ou moins conséquent. Le traitement d'un tel matériau est intrinsèquement difficile, *a fortiori* dans la démarche d'application à de grands corpus qui est la nôtre. De plus ces mentions connexes ajoutent en majorité du bruit à la description de l'événement principal que nous cherchons à capturer. Nous avons par conséquent choisi de ne pas exploiter cette partie dans nos travaux.

Il est important de noter que les dépêches d'agence sont également le fruit de processus rédactionnels spécifiques : lors d'un événement, en particulier s'il est imprévu et important, l'agence émet d'abord une *alerte* ne comportant qu'un titre, puis un titre et un chapeau. Le corps est ensuite développé au fil des minutes et des heures, ajoutant et corrigeant des informations. Ce processus est illustré en figure 1.4 et discuté plus en détail dans la partie 3.1. De plus, une dépêche est un produit vendu à des clients, qui doivent pouvoir l'utiliser à leur convenance, notamment en ayant la possibilité de n'en sélectionner qu'une partie tout en conservant un contenu cohérent.

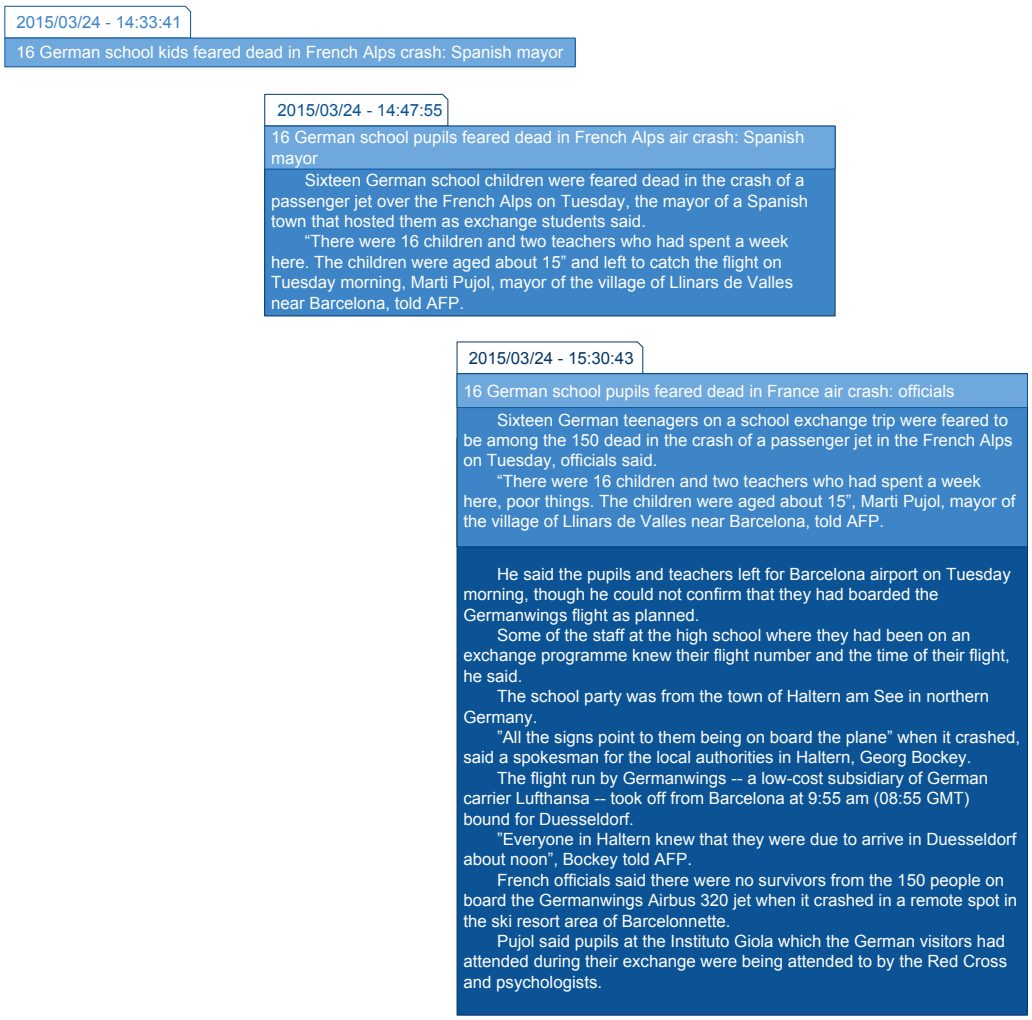


Fig. 1.4: Exemple du processus itératif d’émission des dépêches AFP. La première est une *alerte*, se composant seulement d’un titre. La deuxième ajoute un chapeau et une citation en guise de corps. La dernière développe complètement la nouvelle dans le corps. On peut voir que dans le déroulement de ce processus, le titre et le chapeau ne varient que très légèrement. Sur un fil plus long, ils peuvent rapidement se fixer et être répercutés d’une dépêche à une autre.

1.4.3 Hypothèses de travail

Nous nous fondons sur l’objectif informationnel des chapeaux de dépêches AFP et sur les contraintes fortes régissant leur rédaction, décrites ci-dessus, pour formuler trois hypothèses structurantes pour la suite de nos travaux :

1. le chapeau d’une dépêche AFP peut se réduire à la première phrase de ladite dépêche ;
2. le chapeau d’une dépêche AFP décrit une et une seule instance d’événement ;
3. le chapeau d’une dépêche AFP constitue une description compacte et néanmoins complète de l’instance.

En accord avec ces hypothèses, nous définissons dans la suite de ce manuscrit une dépêche AFP comme l'ensemble textuel formé par son titre et son chapeau, réduit à la première phrase de la dépêche, laissant de côté le corps du document. À cet ensemble textuel s'ajoute la date d'émission de la dépêche, systématiquement fournie en méta-donnée.

Cette décision présente par ailleurs un avantage technique notable, à savoir que restreindre un document à un petit ensemble de phrases indépendantes nous permettra de recourir à des outils de TAL plus fiables et efficaces que ceux exploitant des unités linguistiques plus étendues telles que des suites de phrases coréférentes ou des documents complets, aspect important dans notre perspective de manipulation de grands volumes de données.

1.4.4 Le standard *IPTC Subject Codes*

L'AFP décrit chaque dépêche qu'elle émet par un ensemble de méta-données, parmi lesquelles les *IPTC Subject Codes*, exploités dans ce travail. L'IPTC (*International Press Telecommunications Council*) est un consortium mondial d'agences de presse définissant et maintenant des standards d'échange de données dans le domaine du journalisme. Plus particulièrement, le standard NewsCodes est un vaste vocabulaire standardisé et divisés en différents sous-ensembles, dont les *Subject Codes*, une taxonomie d'environ 1 400 termes, structurée en trois niveaux hiérarchiques[3], conçue pour la qualification de contenus journalistiques.

Nous utilisons ces Subject Codes pour la constitution de notre corpus de dépêches en sélectionnant manuellement, parmi les 1 400 originaux, 72 Subject Codes pouvant être vus comme de nature événementielle, puis en filtrant le corpus des dépêches pour ne conserver que celles portant au moins un de ces Subject Codes. Nous regroupons ensuite ces Subject Codes en 11 "super-catégories" hors de la nomenclature IPTC dans le but de structurer ces Subject Codes choisis en groupes thématiques homogènes. Ce travail fut nécessaire car l'objectif de départ de cette taxonomie n'est pas de caractériser des textes journalistiques dans une perspective événementielle. La liste des Subject Codes retenus ainsi que leur agencement dans la taxonomie IPTC et dans nos propres catégories sont donnés au tableau 1.1.

3. Une visualisation arborescente de cette taxonomie est disponible sur `http://show.newscodes.org/index.html?newscodes=subj`.

1.4.5 Corpus Web

En complément du corpus des dépêches AFP, nous exploitons un corpus d'articles journalistiques collectés sur le Web. Comme pour les dépêches AFP, nous réduisons l'article à sa date de publication, son titre et son chapeau. Bien que les contraintes rédactionnelles liées au chapeau soient plus souples dans les articles de presse que dans les dépêches d'agence, nous n'en retenons, comme dans le cas des dépêches, que la première phrase, en faisant l'hypothèse que la baisse de qualité induite sera compensée par le volume de données considéré. Le tableau 1.2 résume les volumétries des données AFP et Web.

1.4.6 Ressources d'évaluation

Dans la poursuite des objectifs du projet ASRAEL, une direction de recherche indépendante des travaux présentés ici a été explorée, visant à exploiter des données issues de bases de connaissances comme sources de descriptions structurées d'événements (RUDNIK et al., 2019). Ces travaux ont abouti à la constitution de plusieurs ressources qui ont également pu être utilisées pour l'évaluation du présent travail, le recours à des ressources d'évaluation inédites s'étant avéré nécessaire du fait des grands volumes de données manipulés et de l'importante variété événementielle qui y est représentée.

Ces travaux se sont concentrés sur la base Wikidata, une base de connaissances libre et collaborative dont l'objectif principal est la centralisation des données objectives (dates de naissance de personnalités, chiffres d'affaires d'entreprises privées...) utilisées par les projets de la *Wikimedia Foundation*, qui héberge par ailleurs le projet. Ses données sont également ouvertes à tous sous licence *Creative Commons Zero*. Toutes les ressources constituées pour l'évaluation du présent travail mettent en correspondance le nom du fichier associé à une dépêche AFP et l'identifiant d'une structure issue de Wikidata. Ces ressources se divisent en trois niveaux de granularité, un niveau "instance" et deux niveaux "type". Le lecteur trouvera plus de détail sur chaque niveau en parties 3.5, et 4.4, respectivement.

“Super-catégories" thématiques	Niveau de profondeur dans la taxonomie IPTC	
	Deuxième niveau	Troisième niveau
Sport	American football ; Boxing ; Basketball ; Badminton ; Cycling	
Home policy		Impeachment
International relations	Nuclear policy ; Armed conflict ; Migration ; War crime	Extradition ; Immigration ; Economic sanction
Industrial accident	Industrial accident ; Nuclear accident ; Pollution	
Crimes	Police ; Act of terror	Law enforcement ; Theft ; Drug trafficking ; Arson ; Kidnapping ; Investigation ; Bombings ; Assault general ; Gang activity ; Arrest ; Sexual assault
Transport accident	Transport accident	Railway accident ; Maritime accident ; Road accident ; Air and space accident
Natural disasters	Meteorological disaster ; Earthquake ; Fire ; Flood ; Volcanic eruption	Avalanche landslide ; Natural disasters
Elections	Election ; Referenda	National elections ; Poll
Political unrest	Coup d’état ; Massacre ; Riots	Revolutions ; Rebellions ; Political dissent ; Genocide
Social events	Labour dispute ; Strike ; Civil unrest ; Demonstration	
Economic events	New product ; Spin-off	Trade dispute ; Merger, acquisition and takeover ; Privatisation ; Bankruptcy ; Restructuring and recapitalisation ; Board of directors appointment and change ; Contract ; Sales ; Earnings ; Recalls products ; Insider trading ; Nationalisation ; Joint venture ; Layoffs and downsizing

Tab. 1.1: Tableau récapitulatif des 72 Subject Codes (SC) IPTC retenus pour leur nature événementielle. La première colonne décrit les catégories créées manuellement pour grouper les SC de manière plus thématique que la taxonomie IPTC, dont les niveaux hiérarchiques sont représentés dans les deux colonnes suivantes, le premier niveau n’étant pas représenté car trop général.

Année	Nombre de documents	
	AFP	Web
2013	39 974	686 772
2014	38 331	1 218 206
2015	52 724	1 326 037
2016	49 436	777 099

Tab. 1.2: Distribution des volumes de documents traités dans ce travail, année par année, pour les sources AFP et Web.

État de l'art

2

> *C'était quand la dernière fois qu'on s'est retrouvés tous d'accord sur un truc ?*
>
> — **Alexandre Astier**
> Kaamelott, Livre IV, Au service secret de Sa Majesté

2.1 Notion d'événement dans le Traitement Automatique des Langues

L'automatisation du traitement de données événementielles a fait l'objet de multiples tâches et propositions, aux objectifs variés. Plusieurs systèmes de compréhension automatique des années 80 peuvent être identifiés comme des précurseurs du traitement des événements, parmi lesquels on peut citer IPP (*Integrated Partial Parser*) (Lebowitz, 1983) ou GENESIS (R. Mooney, 1985). Portant sur des dépêches d'agence de presse, ils avaient pour vocation de synthétiser les connaissances contenues dans les textes auxquels ils étaient confrontés et modélisaient donc assez naturellement les événements relatés, sous la forme de schémas mis à jour au fur et à mesure que le nombre de documents traités augmente.

Les concepts de *frames* (Minsky, 1974) et de *scripts* (Schank et Abelson, 1977) sont également philosophiquement liés à la notion d'événement. Une *frame* est une structure décrivant une situation, ses participants ainsi que des comportements stéréotypiques ; c'est un concept qui a d'ailleurs structuré les premières campagnes d'évaluation en TAL s'intéressant au traitement des événements, comme nous le verrons un peu plus loin. Le *script* peut quant à lui être envisagé comme une extension de la *frame* prenant en compte l'ordonnancement d'une séquence de situations. Il s'agit d'un concept qui a donné et donne toujours lieu à un grand nombre de travaux exploitant les grandes évolutions techniques du TAL, des méthodes symboliques fondées sur le paradigme *explanation-based learning* (R. Mooney, 1985) aux réseaux de neurones (Pichotta et R. J. Mooney, 2016) en passant entre autres par les modèles probabilistes fondés sur les chaînes de Markov (Orr et al., 2014). Toutefois,

nous avons choisi de ne pas approfondir la direction des *scripts* dans ce travail car nous pensons que l'enjeu majeur de ce type de structure est la modélisation de la composante temporelle, qui n'entre pas en ligne de compte pour nos objectifs.

Le domaine de l'Extraction d'Information est le premier à structurer les travaux consistant à extraire les informations liées à des événements autour de campagnes d'évaluation : d'abord MUC (Message Understanding Conference, GRISHMAN et B. SUNDHEIM (1996)) entre 1987 et 1997, qui s'inscrit en partie dans le cadre de la Phase I du programme TIPSTER de la DARPA et à laquelle succède ACE (Automatic Content Extraction, 1999-2008, DODDINGTON et al. (2004)) puis plus récemment TAC [1] (Text Analysis Conference, 2008-présent), qui prolonge également la piste *Question Answering* de la conférence TREC (Text REtrieval Conference), issue du domaine de la Recherche d'Informations.

Chaque campagne définit une vision de l'événement qui conditionne ses modalités et s'affine à chaque édition : MUC fournit des *templates*, structures constituées par des experts et composées du nom de l'événement auquel s'adjoignent différents arguments (*slots*) caractéristiques de celui-ci. Bien que ces *templates* puissent s'apparenter aux schémas d'événements dont l'induction est l'objet du présent travail (et définis en section 1.3), il est important de noter qu'ils ne sont pas spécifiquement conçus pour la description d'événements mais plutôt autour de la notion de *frame* (B. M. SUNDHEIM et CHINCHOR, 1993) de Minsky, présentée précédemment. On remarquera que bien que ces *templates* constituent des structures génériques, ils sont fournis aux participants en tant que données afin de guider l'extraction des mentions d'intérêt et limitent de fait les campagnes MUC à la description d'instances d'événements telles que nous les avons définies en section 1.3, par l'extraction de réalisations textuelles de granularité plus fine que notre définition de la mention (de l'ordre du terme ou du groupe lexical). Dans la continuité de ce paradigme, chaque nouvelle édition de MUC a augmenté le volume de données mis à disposition des participants ainsi que le nombre et la complexité de ses *templates*. La conférence TAC définira en 2009 la piste KBP (*Knowledge Base Population*) qui reprendra le paradigme de MUC, à la différence que les *templates* ne seront cette fois pas constitués par des experts mais extraits automatiquement à partir d'infoboxes Wikipédia [2] avec notamment une piste dédiée aux événements entre 2014 et 2016 [3].

Prolongeant MUC et reprenant son cadre de travail autour des mentions textuelles, ACE se fixe pour objectif d'identifier et de typer toutes les mentions d'entités dans un corpus de document. La tâche envisagée diffère de celle de MUC en cela qu'il ne s'agit plus d'identifier la mention correspondant à un type cible (le *slot* du *template*) mais

1. https://tac.nist.gov/about/.
2. http://pmcnamee.net/kbp/090601-KBPTaskGuidelines.pdf.
3. https://tac.nist.gov/2014/KBP/Event/index.html.

d'attribuer un type cible (type de l'entité, par exemple `ORGANIZATION`) à l'ensemble des mentions auxquelles il s'applique (DODDINGTON et al., 2004). En cela, ACE peut être envisagée comme une réponse au retour d'expérience qu'a constitué MUC, ce changement de paradigme visant à permettre l'apport d'annotations plus directement exploitables dans le contexte de l'apprentissage supervisé, par opposition à MUC dans laquelle les *templates* ne peuvent constituer qu'une forme de supervision distante.
Les éditions ultérieures d'ACE ajouteront à cette tâche de reconnaissance d'entités la reconnaissance des relations entre différentes entités (par exemple, identifier la relation `SUBSIDIARY` qui lie les mentions d'une maison-mère et sa filiale, deux entités de type `ORGANIZATION`), puis la reconnaissance des événements dans le contexte desquels les entités interagissent (par exemple dans la phrase *she married him*, le terme *married* sera annoté par l'événement `Marry`). L'événement y est cette fois défini comme "une occurrence impliquant des participants, quelque chose se produisant, quelque chose pouvant être décrit comme un changement d'état" (LINGUISTIC DATA CONSORTIUM, 2005).
Le paradigme ACE a eu un grand succès qui a détourné la campagne de ses objectifs initiaux, orientés vers ceux de MUC, pour constituer un terrain de recherche à part se cantonnant à un modèle d'événement atomique, caractérisant et décrivant les interactions d'un déclencheur lexical avec ses arguments.

Toutes ces campagnes fournissent des corpus et des standards d'annotation différents, se voulant de plus en plus pertinents pour les tâches considérées, auxquels s'ajoutera en 2012 le standard ERE (*Entities, Relations, Events*, SONG et al. (2015)), décrit comme une version allégée du standard ACE. Une définition plus précise ainsi qu'une comparaison de ces nombreux standards est effectuée par AGUILAR et al. (2014). Néanmoins, la définition de l'événement, essentiellement celle d'ACE, restera stable.

En parallèle, à partir de 1997 et jusqu'en 2002, le programme *Topic Detection and Tracking* (TDT) est initié par la DARPA, à l'instar de MUC et ACE mais avec des objectifs très différents. En effet, TDT se centre autour de la notion de *topic*, dont la définition évoluera au cours de ses différentes éditions. TDT-1 fait d'abord coïncider *topic* et événement, définissant un *topic* comme "un événement, c'est-à-dire quelque chose se produisant à un instant et un lieu précis et accompagné de toutes les conditions préalables nécessaires ainsi que toutes ses conséquences inévitables"[4] (ALLAN, 2002, Chapitre 2), une définition proche de notre propre définition d'une instance d'événement. Cette définition sera toutefois élargie dès la deuxième édition pour tenir compte des événements connexes découlant de l'événement déclencheur (ALLAN, 2002, Chapitre 2), prenant ainsi une dimension plus thématique. L'autre particularité de TDT est d'introduire la notion de *story*, définie comme "un segment de docu-

4. *"something that happens at some specific time and place along with all necessary conditions and unavoidable consequences."*.

ment thématiquement [5] cohérent incluant au moins deux propositions déclaratives indépendantes portant sur un seul événement" (ALLAN, 2002, Chapitre 2). Cette définition rejoint notre propre définition de la “mention" (section 1.3).
Ces spécificités donnent à TDT une vision de l'événement très différente de celle de MUC et ACE, dépassant le cadre atomique de ces dernières, dont l'influence se ressent dans la nature des tâches proposées, plusieurs d'entre elles étant intrinsèquement non-supervisées ou intégrant des problématiques de cohérence et de séquentialité des événements traités. Le programme a développé cinq tâches (ALLAN, 2002, Chapitre 2) :

Topic Tracking : identifier les mentions se rapportant à un ensemble d'événements connus ;

Link Detection : introduite dans TDT-3 (2000), il s'agit de décider si deux documents traitent du même *topic* ;

Retrospective Event Detection (TDT-1) ou *Topic Detection* (éditions suivantes) : il s'agit de segmenter un corpus de documents en fonction des *topics* qu'ils relatent. Cette tâche est explicitement décrite comme une tâche non-supervisée, en cela qu'aucune information sur la nature des événements ou *topics* recherchés n'est fournie *a priori* ;

First Story Detection : l'objectif de cette tâche est de traiter un corpus de documents en séquence et de décider si le document courant traite du *topic* considéré jusqu'à maintenant ou d'un nouveau. Il s'agit là aussi d'une tâche explicitement non-supervisée ;

Story Segmentation : les systèmes participants doivent traiter une séquence de documents et segmenter ce corpus en *stories*. Cette tâche est envisagée comme une tâche dont la résolution permettra de faciliter les tâches précédentes, les *stories* étant les constituants des *topics*.

TDT peut être vue comme un complément au cadre défini par MUC et ACE, en cela que ses objectifs et les moyens envisagés visent à organiser des informations non-structurées autour de deux axes, thématique et temporel, autour d'un pivot, l'événement, à partir d'une structure latente non-définie. La structuration qui émerge de ce résultat constitue alors la fondation sur laquelle pourront s'appuyer des solutions s'inscrivant dans les objectifs de MUC et ACE. Cette complémentarité est notamment visible dans les travaux de FERRET (1998), dans lesquels une première étape d'identification des domaines thématiques d'un corpus, très proche de la tâche *Topic Detection* de TDT, est complétée par un système structurant les éléments caractéristiques de chaque domaine dans une représentation synthétique assimilable à un schéma d'événement. C'est dans cette dynamique que nous inscrivons le présent travail, en cherchant à faire émerger des structures d'événements sans information préalable sur leur nombre ou leur nature, puis à structurer ces résultats selon une

5. “thématiquement" traduit ici *topically*, c'est-à-dire relatif à un *topic* au sens défini par TDT.

représentation explicite directement issue des *templates* décrits dans MUC. La suite de ce chapitre détaille des contributions partageant ces paradigmes et ces objectifs, tout en nous efforçant de brosser un panorama des tâches et applications centrées sur la notion d'événement dans des acceptions différentes.

2.2 Traitements automatiques de l'événement

La notion d'événement a été exploitée sous différentes formes dans le TAL. Elle sert par exemple de point d'ancrage pour le résumé automatique de documents journalistiques, qu'il s'agisse de générer un chapeau (c'est-à-dire résumer automatiquement un article en une phrase comme le font SUN et al. (2015)), de reconstituer une description répartie sur plusieurs articles (LI et al., 2016) ou de présenter une synthèse d'un grand nombre d'articles traitants d'un sujet spécifié par l'utilisateur via une requête (TOUILEB et DUARTE, 2016).

En analyse de sentiments, établir la polarité d'un événement permet de préciser celle, implicite, des participants les uns par rapports aux autres, comme l'explorent par exemple DENG et WIEBE (2015). Une autre motivation de la détection d'événement dans ce domaine est d'étendre la simple prédiction de polarité, comme dans (GUI et al., 2016) qui identifie les événements à l'origine de la polarité détectée. Dans un esprit semblable, (DING et RILOFF, 2018) défini une catégorisation des causes sous-jacentes à la polarité d'un événement (il peut être positif ou négatif mais agit-il sur la santé, la situation financière... du ou des participants ?).

L'immense majorité des événements se définissant par un contexte spatial et temporel, ces deux dimensions sont particulièrement étudiées. Par exemple, GAO et ZHAO (2018) utilisent des données issues des réseaux sociaux pour prédire la géolocalisation et l'ampleur d'épidémies de grippe et de mouvements sociaux. Il s'agit toutefois d'un champ de recherche moins développé que celui portant sur l'aspect temporel, qui a fait l'objet d'efforts de formalisation beaucoup plus poussés, comme en témoigne l'établissement du standard TimeML et ses dérivés (PUSTEJOVSKY et al., 2010) ainsi que des campagnes TempEval qui l'ont suivi. Historiquement lié au paradigme ACE, TimeML ne constitue pas le seul cadre de travail dans lequel se sont inscrits les travaux sur la temporalité des événements. Un autre domaine de recherche sur le sujet est la reconstitution de chronologies d'événements, soit par le recours à des informations temporelles permettant d'ordonner une séquence d'événements de manière "absolue" comme le proposent NGUYEN et al. (2014) en extrayant les informations temporelles explicites des textes avec l'outil Heideltime (STRÖTGEN et GERTZ, 2012), soit de manière "relative", par la détection des liens (potentiellement implicites) de cause à effet entre des composantes de l'événement (par exemple

en replaçant un séisme avant ses répliques), une approche adoptée par HASHIMOTO et al. (2015) ou KRUENGKRAI et al. (2017). Il faut remarquer que ces deux approches n'ont pas les mêmes objectifs, l'approche “absolue" convenant beaucoup mieux pour la description d'événement au sens de MUC (des instances d'événements tels que ceux que l'on trouvera dans des textes journalistiques) quand l'approche “relative" est plutôt utilisée pour l'ordonnancement d'événements au sens d'ACE, c'est-à-dire de groupes lexicaux composés d'un déclencheur et de ses arguments (se rapprochant dès lors du concept de scripts).

Dans la suite de ce chapitre, nous décrirons plus en détail un certain nombre d'approches d'intérêt pour nos objectifs, se concentrant sur le traitement des événements relatés dans des corpus journalistiques. Nous proposons un cheminement en trois parties correspondant à autant d'échelle de représentation de l'événement, par taille croissante : d'abord des événements d'échelle “atomique" (c'est-à-dire très proche des mentions, nous explicitons cette appellation plus loin), puis à l'échelle de l'instance et enfin celle de la structure générique (agrégeant plusieurs instances). Cette construction reflète également l'articulation de notre propre travail, qui va de la mention à l'instance et de l'instance au schéma (via une étape d'identification des types d'événements, par laquelle nous atteignons l'échelle agrégeant plusieurs instances).

2.2.1 Traitement d'événements atomiques

La première échelle de représentation de l'événement sur laquelle nous nous attardons est aussi la plus fine, restreignant l'événement à une ou quelques mentions textuelles, généralement sous la forme d'un déclencheur lexical et de ses arguments syntaxiques. Il s'agit à notre avis de la façon minimale de représenter un événement, d'où la dénomination “événement atomique". Cette vision fine, fortement liée aux mentions textuelles, s'est développée autour du paradigme des *frames* [6], notamment par le biais des campagnes ACE. Nous intégrons également dans cette section des travaux centrés autour de la modélisation de *scripts* car nous nous concentrons ici sur l'échelle adoptée pour la représentation des événements (à savoir, dans le cas des travaux exposés, cette échelle fine, atomique) et considérons la structuration de ces représentations selon l'axe temporel (caractéristique des *scripts*) comme orthogonal. Par ailleurs, si les premières propositions présentées se cantonnent à des corpus de faible volume, souvent avec des systèmes fondés sur de l'apprentissage supervisé cherchant à tirer pleinement parti des annotations disponibles, plusieurs travaux vont se donner par la suite pour objectif d'étendre ce paradigme aux grands volumes de données en domaine ouvert, en ne se reposant plus sur des ensembles de relations connues à l'avance mais en les inférant à partir des données.

6. Bien qu'aucune limite d'échelle ne soit intrinsèquement posée par celui-ci.

SEKINE (2006) propose un système non supervisé en plusieurs étapes. D'abord, un sous-ensemble de documents est sélectionné de manière classique à partir d'un corpus global sur la base d'une requête formulée par un utilisateur. Ces documents sont ensuite analysés pour en extraire des arbres de dépendances et identifier les plus fréquents d'entre eux dans le sous-ensemble sélectionné par rapport à l'ensemble du corpus (selon le même principe que la pondération tf.idf). Les participants de ces sous-arbres de dépendances sont ensuite rendus génériques pour former les relations d'intérêt. Des équivalences entre ces relations sont alors identifiées à l'aide d'une base de paraphrase construite en amont. Enfin, ces relations sont ré-appliquées aux documents du sous-ensemble pour en extraire les mentions particulières (non génériques) qui seront regroupées en "table" de participants équivalents formant la réponse à la requête de l'utilisateur. Le résultat est un système de Recherche d'Information orienté événement dans lequel l'utilisateur renseigne un ou plusieurs déclencheurs événementiels de son choix et se voit retourner les participants génériques et particuliers (mentions dans le corpus) associés.

En parallèle, SHINYAMA et SEKINE (2006) adoptent une approche différente, toujours fondée sur l'extraction de relations entre entités. Les auteurs commencent par effectuer un clustering de documents issus de sources différentes pour dégager des instances d'événements. Au sein de chaque instance ainsi identifiée, des relations sont extraites sous la forme d'arbres de dépendances dont les entités sont ensuite connectées aux occurrences dans les autres documents et normalisées pour ne conserver qu'une forme de triplet {sujet, prédicat, objet}. Enfin, ces arbres font l'objet d'un nouveau clustering par agrégation de ceux communs à plusieurs clusters d'instances. Sont ainsi obtenus des ensembles de triplets décrivant des relations équivalentes. Contrairement à l'approche précédente, une composante événementielle de plus haut niveau est ici explicitement intégrée mais le but et les résultats obtenus n'ont rien à voir avec des éléments susceptibles de caractériser des événements au sens large ; la dernière étape de clustering ne vise pas à identifier des types d'événements ou des ensembles de relations complémentaires mais plutôt des classes d'équivalence entre relations décrivant des événements atomiques, se ramenant au cadre défini dans ACE. Ces deux travaux continuent néanmoins d'exploiter des volumes de données modestes (bien que très supérieurs à ceux d'ACE) issus de sources journalistiques.

D'autres travaux se reposent sur des approches semi-supervisées (au contraire des deux précédentes qui étaient non supervisées). C'est le cas par exemple du système présenté dans (BANKO et al., 2007) et qui introduit l'*Open Information Extraction* (OIE), un paradigme d'Extraction d'Information applicable aux grands volumes de données hétérogènes du Web sous la forme de triplets {sujet, prédicat, objet}. Ce système se divise en trois modules, dont le premier extrait des triplets candidats sans supervision, lesquels sont filtrés par le deuxième module consistant en un classifieur entraîné de manière supervisée pour décider de la pertinence de chaque

triplet qui lui est présenté. Enfin, le dernier module normalise les triplets extraits et estime la probabilité que chaque triplet non normalisé soit une instance de sa version normalisée afin de les ordonner. Cette approche présente l'avantage de s'appliquer aux enjeux du Web, même si le nombre très important de triplets récupérés pose des questions sur les cas d'usages directs de cette méthode. Néanmoins sa capacité à faire émerger un grand nombre de triplets décrivant des relations variées (et de fait pas toujours de nature événementielle), à partir de grands volumes de données hétérogènes, fera de l'OIE un paradigme de référence, réutilisé et étendu dans beaucoup de travaux tels que (CHRISTENSEN et al., 2010) qui complète l'approche du système par l'exploitation des résultats de l'analyse morpho-syntaxique du texte traité, (DEL CORRO et GEMULLA, 2013), qui introduit les informations issues d'un système d'étiquetage des rôles sémantiques pour affiner l'identification des triplets ou encore (ROY et al., 2019), qui utilise des techniques d'*ensemble learning* pour tirer parti des spécificités de trois systèmes d'OIE distincts pour produire de nouvelles extractions.

CHAMBERS et JURAFSKY (2008) s'éloignent de ces représentations pour aborder la description d'événements sous l'angle du *script* tel que formalisé par SCHANK et ABELSON (1977), c'est-à-dire de séquences ordonnées d'actions impliquant des participants, que les auteurs appellent *narrative (event) chains*. Ce travail puise ses inspirations dans des travaux sur la résolution de coréférence. Une *narrative chain* est structurée autour d'un *protagonist*, c'est-à-dire un participant apparaissant dans chaque situation de la *narrative chain*, selon une hypothèse de cohérence narrative : si différents verbes impliquent le ou les mêmes participants, alors ces verbes narrent un même événement. Plus concrètement, ces structures sont construites par un système semi-supervisé en trois temps.
En premier lieu, une matrice de cooccurrence des événements est construite par le calcul de l'information mutuelle (*Point-wise Mutual Information*, abrégée PMI dans la suite de ce document) sur les cooccurrences des arguments verbaux. Le chaînage des événements est alors amorcé en groupant les événements d'un même document dans une même chaîne, puis en attribuant chaque événement restant du corpus à la chaîne dont l'un des éléments maximise la PMI.
Dans un deuxième temps, les événements de chaque chaîne ainsi constituée sont ordonnancés temporellement. Les auteurs recourent pour cela à deux étapes de classification supervisée : d'abord, les événements sont étiquetés selon leurs attributs temporels grammaticaux (temps, aspect…) par un modèle SVM (*Support Vector Machine*) fondé sur des traits linguistiques, puis ces informations sont utilisées (parmi d'autres) par un second SVM pour produire une relation temporelle d'antériorité entre les événements deux à deux. Cet ordonnancement permet finalement de consolider des *scripts* par la production de graphes orientés dont les nœuds sont formés par les événements et les arêtes indiquent ces relations temporelles.
Ces structures sont explicitement envisagées par les auteurs comme extensibles,

ce qui les distinguent des *scripts* dont elles sont inspirées, qui ont pour leur part vocation à être statiques et auto-suffisants. Ainsi, l'ordonnancement des événements dans une *narrative chain* n'est pas vu comme définitif mais simplement relatif aux événements qui la constituent à un instant t.

Ce concept de *narrative chain* est étendu dans (CHAMBERS et JURAFSKY, 2009) dans lequel les auteurs visent à produire des regroupements cohérents de *narrative chain* baptisés *narrative schemas*, ceci en dépassant la focalisation des *narrative chains* autour d'un seul participant (le *protagonist*). Dans cette optique, les auteurs incorporent des informations contextuelles afin de compléter la vision purement statistique du modèle de base. Pour ce faire, ils commencent par typer le participant récurrent de la chaîne (le *protagonist*), puis enrichissent la construction des chaînes d'événements en ajoutant à la compatibilité statistique de ces dernières une information sur le type de tous les autres participants afin de renforcer la cohérence narrative des chaînes induites. La construction des *narrative schemas* est alors effectuée selon une généralisation de l'algorithme de construction des *narrative chains*, en ajoutant chaque chaîne au schéma dans lequel tous ses arguments existent au préalable.

La représentation automatique d'événements atomique a également attiré l'attention de chercheurs issus du domaine du Web sémantique, avec plusieurs projets de grande envergure. C'est le cas par exemple du projet XLike, dont les composantes de TAL sont présentées dans (PADRÓ et al., 2014) et qui se donne pour objectif de décrire et d'organiser du contenu journalistique à partir de *frames* en s'appuyant sur un pipeline classique d'outils de TAL (analyse morpho-syntaxique, reconnaissance d'entités nommées, etc.). Si le projet reste au niveau de la *frame*, c'est en partie car son objectif est un traitement multilingue de l'information : il exploite en effet des *pipelines* d'analyse pour 7 langues, ce qui aboutit à une architecture complexe, même pour l'extraction de *frames*.

Bien que l'un des plus représentés en TAL, ce paradigme d'extraction d'événements "atomiques" trouve rapidement ses limites lorsqu'on le confronte aux besoins d'autres disciplines, comme le font observer SPRUGNOLI et TONELLI (2017) pour le cas de l'étude historique ou, comme on pourra le constater dans la partie suivante, pour celui du milieu journalistique.

2.2.2 Traitement d'instances d'événements

Une autre vision de l'événement ayant donné lieu à de nombreux travaux est celle de la représentation d'instances, c'est-à-dire d'occurrences dans un contexte spatial et temporel particulier d'un changement remarquable du réel. Ce paradigme rejoint plutôt celui des conférences TDT et les travaux réalisés l'ont été en grande partie

afin d'exploiter et de faciliter l'accès aux grands volumes d'informations du Web, enjeu qui est devenu de plus en plus prégnant au fil du temps.

Dans cette perspective, la plupart des approches prennent la forme d'agrégateurs d'informations. Par exemple, AZZOPARDI et STAFF (2012) exploitent différentes sources journalistiques du Web et se fonde sur une version non-paramétrique de l'algorithme K-means (le nombre de clusters n'est pas fourni à l'avance à l'algorithme) fonctionnant "en ligne", c'est-à-dire que l'algorithme reçoit chaque document à traiter l'un après l'autre, en un flux continu. Le processus est amorcé à partir d'un petit corpus dont le traitement fournit les premiers clusters, puis l'algorithme est déployé dans sa version en ligne. Si la similarité entre un document entrant et tous les clusters existants ne dépasse pas un certain seuil, ce nouveau document devient un centroïde (c'est-à-dire qu'une nouvelle instance est considérée). Afin de limiter les imprécisions de clustering à long terme (c'est-à-dire que plus il y aura de documents traités, plus la similarité intra-classe des clusters va baisser) et de garantir la capacité de ce système à passer à l'échelle (par exemple pour ne pas tomber dans le cas où un document serait comparé à un nombre infini de clusters), les clusters n'ayant pas reçus de nouveaux documents pendant une durée donnée sont "gelés" : ils ne feront plus partie des clusters dont on comparera la similarité à celle des nouveaux documents, considérant que les instances qu'ils décrivent est terminée. Ce système fait donc émerger les structures événementielles d'un flux continu de documents afin de clarifier la navigation de l'utilisateur.

Suivant une direction plus structurée et proche de celle favorisée par TDT, GLAVAŠ et ŠNAJDER (2013) proposent la description d'instance selon un modèle proche de celui des 5W, par l'extraction de mentions d'intérêt dans les documents d'un corpus, à savoir celles situant le lieu, la temporalité, le déclencheur événementiel, son agent et son patient. Ces mentions résumant le document sont structurées en un graphe comparé à tous les autres graphes de tous les autres documents par un *graph kernel* (BORGWARDT, 2007), une alternative plus efficace aux techniques de comparaison de graphes classiques visant à mesurer la similarité entre des graphes.

Dans la lignée des agrégateurs de documents, CONRAD et BENDER (2016) accordent une place particulière à la dimension temporelle, qui sert de point d'entrée à leur système de recherche d'information fondé sur la description d'instances d'événements à partir de NewsRoom, un dépôt de documents journalistiques multi-source géré par l'agence Thomson-Reuters. Étant donnée une fourchette temporelle, un sous-ensemble de documents est sélectionné dans NewsRoom. Une première phase de clustering hiérarchique est appliquée avec pour objectif de découper ce corpus en groupes de forte similarité (doublons ou quasi-doublons). Ces clusters hautement similaires sont ensuite utilisés comme amorces pour une seconde phase de clustering hiérarchique visant à étendre la couverture des événements à décrire et dont le

critère de regroupement combine la similarité de deux vecteurs, l'un construit à partir des termes du document, l'autre à partir d'entités nommées identifiées dans le document.
Cette approche présente l'inconvénient d'être moins générique que la précédente, la dernière phase de clustering étant amorcée par l'exploitation de la *slugline* (une dénomination standardisée de l'instance propre à l'agence Thomson-Reuters) associée à chaque dépêche présente dans les clusters de l'étape précédente. Bien que les détails de l'utilisation de cette information restent flous, elle semble constituer une limitation notable pour deux raisons : premièrement, elle suppose que des clusters ne contenant pas de dépêches Reuters ne seront pas traités, ce qui limite l'intérêt d'un recours à un corpus multi-source. Ensuite, ces *sluglines* sont certes normalisées mais uniquement au sein de l'agence Reuters. Si la sélection de ces *sluglines* est importante dans le fonctionnement du système, alors l'utiliser sans dépêches Reuters peut s'avérer difficile (en demandant un effort d'adaptation) voire impossible (si on ne dispose pas d'une méta-donnée comparable).

Afin de permettre une plus grande flexibilité dans les angles d'analyse offerts à l'utilisateur, Y. CHEN et al. (2017) s'orientent quant à eux vers la construction de graphes d'événements. Leur approche est constituée de trois grandes étapes.
La première consiste à identifier les instances d'événements, d'abord en découpant le corpus en blocs de documents, un bloc représentant une période temporelle de cinq mois. Puis un clustering fondé sur l'approche LDA (*Latent Dirichlet Allocation*) (BLEI et al., 2003) est appliqué à chaque bloc, afin de produire 25 *topics* sous la forme de clusters de termes. Les documents du bloc temporel sont alors regroupés en fonction de la distance entre leur contenu textuel et celui constitué par ce vocabulaire, utilisé comme centroïde, pour constituer des clusters correspondant approximativement à des instances d'événements. Ce traitement (clustering LDA en 25 *topics* puis regroupement des documents par similarité entre leur contenu et le vocabulaire de ces *topics*) est ensuite ré-appliqué à chaque instance afin d'en dégager les "sous-événements"[7]. Par ailleurs, le corpus étant découpé en blocs temporels traités de manière indépendante, une étape de réconciliation est mise en place pour reconstituer les instances décrites par des documents à travers plusieurs blocs temporels adjacents.
La deuxième étape vise à construire des graphes représentant chaque instance identifiée lors de l'étape précédente, à partir des entités nommées (restreintes aux personnes, organisations et lieux) et des relations (au sens d'ACE, ici les relations de type `General-Affiliation`, `Personal-Social`, `Part-Whole`, `Physical`, `Organization-Affiliation`) qui les relient. Les auteurs recourent pour cela à des classifieurs supervisés.
La dernière étape consiste en une analyse des résultats par le recours à un logiciel de

7. D'après la terminologie des auteurs.

visualisation dans le but d'illustrer la flexibilité de l'approche. Elle ne contribue donc pas à raffiner la description des événements inférés et nous ne nous attarderons donc pas sur sa description.

Parmi les grands projets de Web sémantique s'étant penchés sur le traitement des événements, le projet NewsReader (ROSPOCHER et al., 2016) fait partie de ceux qui ont porté leurs efforts sur la description d'instances. Comme le projet XLike, il s'appuie sur un pipeline d'analyse de TAL classique dans différentes langues pour produire des descriptions génériques d'instances d'événements suivant le Simple Event Model, dérivant lui-même des 4W factuels du modèle 5W, c'est-à-dire le *Who* (entités participant à l'événement), *What* (l'action de l'événement), *When* (l'ancrage temporel) et *Where* (l'ancrage spatial). Toutes les mentions et les liens identifiés entre elles sont normalisés *via* des bases de connaissances afin d'en produire une représentation standardisée, qui est ensuite traitée afin d'être visualisée et analysée par les utilisateurs.

Dans le prolongement de cette idée de représentation standardisée, (NANNI et al., 2017) constitue moins un travail d'induction de structures à proprement parler qu'une tentative de consolider la description d'instances d'événements à travers différentes sources. Partant de la dénomination normalisée d'une instance d'événement (ici un *item* DBpédia), le système présenté récupère un ensemble de pointeurs dans DBpédia vers l'événement en question. En est extrait un ensemble d'entités caractéristiques, à partir de la page Wikipédia de l'événement. Ces entités sont ensuite classées par ordre de pertinence vis-à-vis de l'événement par comparaison à des plongements RDF2Vec associés (un plongement donné étant fondé sur le graphe RDF des liens entre éléments de la base) puis filtrés pour n'en conserver qu'un certain nombre. Leurs contextes d'apparition (à l'échelle phrastique) sont ensuite récupérés pour en extraire un plongement neuronal. Le nom de l'événement recherché est alors projeté dans le même espace pour étendre les résultats récupérés. Enfin, les documents résultant de cette recherche sont reclassés par ordre de pertinence avant d'être présentés.

On constatera que ces travaux se sont focalisés sur les données journalistiques en grand volume avec pour objectif de les organiser et de rendre leur contenu et leur variété exploitable plus facilement pour leurs utilisateurs. Il s'agit donc d'un domaine qui a bénéficié à la fois d'un cadre applicatif favorable et de travaux antérieurs structurants, ce qui explique sans doute qu'il soit celui ayant le plus de systèmes "aboutis" (dans le sens où ils peuvent être utilisés directement par un utilisateur pour satisfaire un besoin informationnel). Dans la partie suivante, nous discuterons des travaux visant à extraire des informations de ces grands volumes de données non plus seulement dans le but de l'organiser mais également d'en synthétiser la substance.

2.2.3 Traitement de structures génériques d'événements

L'utilité de l'induction de structures génériques d'événements (sous la forme de schémas mais pas exclusivement) est reconnue depuis longtemps. Ce domaine reste toutefois moins exploré que ceux des événements atomiques et des instances, en raison de la difficulté à expliciter les composantes latentes qui constituent la substance d'un événement à travers ses différentes réalisations. Un certain nombre de travaux d'intérêt ont malgré tout été menés, à commencer par (FILATOVA et al., 2006), qui se propose d'induire des *domain templates* reprenant la structure des *templates* de MUC, un *domain* s'apparentant à un type d'événement. Ce cadre est assimilable à notre concept de schéma d'événements. Les auteurs se donnent pour objectif d'étendre la couverture des types d'événements décrits par les structures manuelles constituées pour MUC à l'aide d'un outil entièrement automatique. Ils présentent pour ce faire un système statistique d'extraction d'information travaillant sur un corpus de documents segmenté selon les types d'événements à décrire et fondé sur la caractérisation des instances par les déclencheurs verbaux propres à chaque type d'événement. Cette liste de verbes est ensuite filtrée et utilisée pour analyser le corpus et identifier les groupes lexicaux (incluant potentiellement plusieurs propositions) les plus fréquents sous la forme de leurs arbres syntaxiques. Les instances ainsi décrites sont groupées afin de faire émerger les arguments (*slots*) des types d'événements par la fusion des arbres syntaxiques récurrents et en ne conservant que ceux qui, après fusion, sont représentés au moins deux fois dans chaque instance. Enfin, le schéma est consolidé en agrégeant tous les arbres partageant le même verbe et en retirant les informations syntaxiques afin de ne conserver que le verbe, son ou ses dépendants et leurs étiquettes (morpho-syntaxique ou d'entité nommée). Chacune de ces structures forme un argument du schéma. Le schéma est par conséquent constitué de l'ensemble de tous les arguments trouvés pour un type d'événement donné.

Cette approche est toutefois limitée par la nécessité de connaître à l'avance les types des *templates* que l'on souhaite modéliser afin de segmenter correctement le corpus en amont du traitement. Le travail de constitution des *templates* est donc automatisé mais requiert tout de même une connaissance du corpus et une expertise humaine non-négligeable. De plus, cette segmentation est indispensable et doit rester relativement précise car sans elle, la phase de fusion des arbres syntaxiques pourrait être perturbée pour des arguments proches ou communs (par exemple, les victimes dans le cas d'une catastrophe naturelle ou d'un accident). De même, cette segmentation est utilisée pour identifier les déclencheurs verbaux caractéristiques de chaque type à décrire selon un principe de tf.idf. Or on sait que beaucoup de types d'événements peuvent partager des caractéristiques communes (les plus évidentes étant la date et la localisation), dont l'identification est plus difficile par cette méthode puisque les déclencheurs associés se verront mécaniquement attribuer des scores faibles. Enfin, l'identification des événements est faite uniquement sur la base de déclen-

cheurs verbaux, ce qui exclut les événements décrits par des nominalisations, forme d'expression événementielle pourtant non-négligeable ((TANNIER et al., 2012) et (TANNIER, 2014), page 22).

Dans ses travaux, Chambers s'attelle également par deux fois à l'induction non supervisée de schéma d'événement en adoptant deux approches radicalement différentes. Celle de (CHAMBERS et JURAFSKY, 2011) vise à délaisser la notion d'ordonnancement temporel propre aux *scripts* et envisage le schéma comme un *template* plus proche de la forme proposée dans MUC, caractérisé par un ensemble d'événements caractéristiques sous la forme de déclencheurs lexicaux (verbaux et nominaux) d'une part et de rôles sémantiques associés aux événements d'autre part.
L'identification des déclencheurs lexicaux constitue la première étape du système, afin d'identifier les types d'événements dans le corpus, et s'effectue par l'application d'un clustering hiérarchique fondé sur une variante de la PMI prenant en compte la distance textuelle entre les déclencheurs identifiés. Ces ensembles caractérisant des types d'événements sont induits sur le corpus MUC-4, dont la taille très restreinte constitue un obstacle à l'induction de schémas riches. C'est pourquoi les auteurs utilisent ces structures induites comme amorce et les appliquent à un corpus plus large afin d'augmenter la couverture lexicale de leurs proto-schémas.
Enfin, ces ensembles de déclencheurs lexicaux forment la base à partir de laquelle les arguments du schéma sont extraits, par regroupement des relations syntaxiques identifiées entre les termes. Pour favoriser l'induction d'arguments se rattachant à l'événement décrit et non à des rôles universels (patient/agent) ou spécifiquement liés aux termes regroupés, les auteurs proposent une mesure de similarité des dépendances syntaxiques combinant un aspect temporel exploitant des chaînes de coréférence pour modéliser l'ordonnancement des événements et un aspect paradigmatique fondé sur les arguments syntaxiques des déclencheurs considérés. Cette mesure est ensuite utilisée pour produire un clustering hiérarchique des relations syntaxiques de chaque type, produisant les arguments du type et complétant la construction du schéma.
Néanmoins, l'objectif de s'abstraire de la notion d'ordonnancement propre aux *scripts*, non nécessaire pour l'induction de schémas, n'est qu'en partie atteint. Cette composante temporelle n'est pas recherchée ou modélisée explicitement par un module et ne constitue donc pas un enjeu de modélisation mais elle est toujours utilisée en support de la construction des arguments des schémas par l'exploitation des chaînes de coréférence. Ce recours à la temporalité implicite du texte est d'autant plus limitant que la résolution de coréférence reste, même aujourd'hui, une tâche difficile que les outils disponibles implémentent avec des performances variables.

CHEUNG et al. (2013) proposent d'induire des *frames*, en rapprochant néanmoins leur définition de celle des *scripts* de Schank. S'éloignant comme les travaux de Chambers de la vision participants/relations propre à ACE, les auteurs présentent un

modèle génératif non-paramétrique fondé sur les propositions syntaxiques (groupes verbaux et nominaux) d'un document, capturant à la fois l'émission de ces dernières (c'est-à-dire qu'elles apparaissent dans le document) et les transitions entre elles. Chaque proposition est modélisée par un ensemble de variables latentes, la première d'entre elles étant une variable binaire déterminant si la proposition considérée sera de type "contenu" ou "contexte". Cette variable a pour but de permettre aux informations redondantes entre différents événements d'être capturées par un ensemble de *frames* différent de celui dédié à la caractérisation de chaque événement, afin que ce dernier soit moins bruité. Cette première décision est importante car elle conditionne à la fois le choix de la *frame* associée à la proposition, laquelle détermine l'événement décrit, mais également le choix du déclencheur lexical de l'événement (observable du modèle, typiquement un nom ou un verbe) émis, en conjonction avec la variable d'événement. Cette dernière influence quant à elle le choix du *slot* associé, qui émettra à son tour les deux autres types d'observables de la proposition, à savoir les arguments du déclencheur (dépendants syntaxiques) et leurs dépendances syntaxiques.
Cette modélisation de la structure interne d'une proposition garantie que tous les mots d'une même proposition se verront attribuer la même *frame* mais pas forcément le même événement. Néanmoins, les auteurs souhaitent aussi modéliser deux phénomènes de transition : ils appliquent pour ce faire l'hypothèse markovienne à la transition d'une proposition à une autre, en modélisant le changement de *frame* en fonction du précédent. Cette hypothèse est reprise pour la transition d'un événement à un autre au sein d'une même *frame* (et donc d'une même proposition). Ces phénomènes de transition sont pondérés par un paramètre appelé *stickiness factor*, agissant comme une forme de régularisation encourageant une certaine cohérence entre les événements au sein des *frames* et entre les *frames* de propositions voisines. En conséquence, les états des variables latentes associées à la *frame* et à l'événement d'une proposition sont également conditionnés par une variable latente modélisant le phénomène de transition associé.
Nous avons respecté dans cette description la terminologie des auteurs, néanmoins le travail qu'ils présentent est évalué en suivant le protocole de (CHAMBERS et JURAFSKY, 2011) et on peut constater que les structures évaluées tiennent plus de la représentation synthétique hybride que de la *frame* au sens le plus largement utilisé dans la littérature (et décrit ici comme une représentation "atomique").

Suivant une approche similaire, CHAMBERS (2013) abandonne l'approche *pipeline* qui caractérisait ses précédents travaux pour adopter également un modèle génératif. Cette approche présente notamment l'avantage d'être intégrée (*end-to-end*), réduisant les problèmes de propagation d'erreur liés aux approches *pipeline*. Le modèle observe des entités prenant la forme de triplets composés d'une tête syntaxique, de l'ensemble de ses dépendants et des relations qui les relient ainsi que d'un ensemble complémentaire d'entités, typiquement des entités nommées.

Un document est alors modélisé comme un mélange d'événements. Pour chaque entité, un schéma de rattachement est émis selon ce mélange, duquel sont sélectionnés un *slot*[8] et un prédicat de l'entité. Le *slot* sélectionné permet quant à lui d'émettre la tête de l'entité, les entités nommées qui lui sont associées ainsi que les dépendances syntaxiques entre les prédicats et la tête. Le choix de ne pas faire dépendre le prédicat de la variable latente associée au schéma est fait pour permettre à cette dernière de ne pas avoir à modéliser la syntaxe (reportée au niveau de la variable *slot*) et pour que le choix du prédicat conditionne les autres sélections.
Comme beaucoup de modèles génératifs dans ce domaine, l'une des principales limitations de celui-ci est le fait qu'il soit paramétrique, requérant à la fois que le nombre de *templates* et de *slots* soit fixé *a priori*. Il est également entraîné et évalué sur les données de MUC, peu variées et peu volumineuses (mais devenues au fil du temps le standard des protocoles d'évaluation). Une expérience supplémentaire visant à montrer la capacité de ce modèle à apprendre sur un volume de données plus faible (à paramètres fixes) que l'approche précédente rapporte des performances moindres, ce qui suggère une sensibilité non négligeable de ce paramétrage.

Le domaine du Web sémantique a également produit des travaux dans le domaine. Par exemple, dans (KUZEY et al., 2014), la caractérisation de types d'événements vise à peupler une base de connaissances et se déroule en trois étapes. La première attribue un type d'événement normalisé à chaque article considéré en faisant le lien entre l'article et une catégorie d'événement Wikipédia, puis entre cette catégorie et un type d'événement dans WordNet, qui sera l'étiquette finale. Puis cette étiquette est utilisée entre autres informations pour grouper les articles selon deux axes, thématique et temporel. À cette fin un graphe est construit, dans lequel un sommet représente un article et une arête peut figurer soit une relation entre des entités des deux articles qu'elle connecte (auquel cas elle n'est pas orientée), soit une relation d'antécédence chronologique entre les articles (auquel cas elle est orientée). Ces deux dimensions (thématique et temporelle) peuvent ainsi être optimisées conjointement par une stratégie de regroupement fondée sur la réduction de la granularité du graphe (*graph-coarsening*). Le résultat de ce système est un ensemble de documents, d'entités, de classes sémantiques et de dates en réponse à une requête.
Cette approche présente intuitivement la faiblesse de son couplage à Wikipédia : en exploitant les données Wikinews et en standardisant le type d'événement associé en passant par Wikipédia dans leur évaluation, la possibilité de cas défavorables semble intuitivement réduite mais si un corpus différent de Wikinews est utilisé, son recouvrement avec les événements Wikipédia paraît moins évident.

Dans la continuité de l'approche de CHAMBERS (2013), NGUYEN et al. (2015) proposent un modèle génératif intégrant les modifieurs des têtes lexicales afin d'amé-

8. On utilisera *slot* pour éviter l'ambiguïté avec le terme argument.

liorer la désambiguïsation des participants aux événements identifiés. Le modèle élimine également la variable latente capturant la distribution de schémas d'événements attribuée au document, ce qui permet au modèle de ne plus dépendre de deux hyper-paramètres (le nombre de schémas et le nombre de *slots* par schéma) mais d'un seul (le nombre de *slots*). Néanmoins cette modification empêche aussi le modèle de structurer les arguments appris en schémas, ce qui entraîne donc le recours à un post-traitement reposant sur des heuristiques de mise en correspondance des arguments inférés avec ceux de ressources comme MUC, afin de grouper les rôles autour de types.

Par ailleurs, ce modèle est utilisé comme base de travail dans (LIU et al., 2019). Partant d'instances sous la forme de groupes d'articles de presse, le modèle est appliqué pour en induire des arguments en conservant la forme de surface de la tête lexicale de chaque entité mais en remplaçant celles des autres informations par un plongement neuronal pré-calculé. Le modèle introduit également une variable latente capturant le type d'événement associé à l'instance considérée pour contraindre la structuration des arguments autour de types et faciliter l'émergence de schéma. En cela, les auteurs font le chemin inverse de l'étude de NGUYEN et al. (2015) pour se rapprocher du modèle de CHAMBERS (2013), à ceci près que leur variation associe un seul type d'événement à chaque instance (groupant plusieurs documents) quand CHAMBERS (2013) modélise chaque document comme un mélange de plusieurs types.

Sortant de cette ligne de recherche fondée sur les modèles génératifs, SHA et al. (2016) proposent l'apprentissage de schémas selon une approche centrée sur des entités de la forme classique (tête lexicale, dépendant syntaxique, type de la dépendance), à laquelle les auteurs ajoutent l'hyperonyme de la tête afin de favoriser le regroupement entre les instances et donc la couverture des schémas. Guidé par l'hypothèse que les entités d'un même document appartiennent au même schéma, le modèle représente la cohérence des arguments au sein d'un schéma en combinant la PMI des entités d'un même document et la similarité cosinus des plongements lexicaux Word2Vec des têtes d'entités entre les documents, se donnant par là même un degré de liberté supplémentaire par rapport au modèle de CHAMBERS (2013). Les auteurs modélisent ensuite l'attribution de mentions à des arguments du schéma par un score de similarité sommant la similarité cosinus des dépendants de la tête de chaque entité et celles des hyperonymes desdites têtes.
Ces mesures permettent respectivement l'apprentissage de la composition de chaque schéma (en termes d'arguments) et de celle de chaque argument (comme un regroupement de mentions). Cet apprentissage est réalisé de manière jointe, avec deux contraintes supplémentaires : la minimisation du nombre de schémas attribués à une même phrase dans un document et la maximisation du nombre d'arguments attribués à une même phrase.

Comme ses prédécesseurs, ce modèle est paramétrique et il est intéressant de remarquer que le nombre de *templates* et de *slots* fixés par les auteurs après optimisation correspond exactement à ceux du corpus d'évaluation, ce qui est inhabituel car les modèles génératifs fonctionnent généralement mieux avec des valeurs supérieures à celles recherchées afin de laisser de la latitude au modèle, au prix de clusters de moindre qualité qu'il faudra ensuite filtrer.

L'induction de schémas par apprentissage conjoint des arguments et des types qui les regroupent est également l'approche adoptée par HUANG et al. (2016). Pour ce faire, les auteurs calculent d'abord deux types de plongements lexicaux.
Le premier a pour objectif de projeter les termes du corpus d'intérêt dans un espace sémantique commun. Pour cela, chaque terme est désambiguïsé et normalisé dans Wordnet et la forme résultante est identifiée dans OntoNotes. Des plongements lexicaux sont appris (selon le modèle Skip-gram) sur ces représentations normalisées et désambiguïsées, permettant la construction de plongements moins polysémiques que le Skip-gram classique.
Le second type de plongement vise à compléter le premier par l'ajout d'informations contextuelles et se fonde sur l'identification dans le corpus des déclencheurs lexicaux d'événements et des dépendants syntaxiques de ces déclencheurs. Ces structures de termes dépendants sont ensuite rendues plus génériques par la substitution de formes normalisées (issues d'OntoNotes et FrameNet) aux formes de surface. Ces plongements sont appris par un auto-encodeur.
Ces deux représentations sont ensuite exploitées pour construire deux mesures de similarité ; l'une pour les déclencheurs et l'autre pour les arguments. Un clustering spectral est ensuite appliqué sur les déclencheurs et les arguments pour dégager des types sous la forme de clusters de paires déclencheur/argument. Afin que ces deux phases de clustering (sur des ensembles distincts) se fasse de manière conjointe, chacune de ces mesures intègre une contrainte de cohérence, c'est-à-dire que le clustering des déclencheurs favorisera le regroupement de ceux partageant des arguments proches et vice-versa.
Ce travail est l'un des rares à intégrer une heuristique de nommage automatique des types et des arguments des événements obtenus en attribuant le déclencheur le plus proche du centroïde du cluster comme nom du type associé et en attribuant aux arguments leurs rôles dans différentes ressources (FrameNet ou VerbNet ou PropBank successivement si la ou les précédentes ne donnent pas de résultats). C'est un ajout particulièrement bienvenu considérant que les structures présentées prennent la forme de collections de *frames*, un format difficile à interpréter en tant que tel comparé au modèle argument/valeurs (*slot/slot-filler*) fixé par MUC.
L'une des contributions de ce travail est de tirer parti de la désambiguïsation sémantique pour l'apprentissage de plongements lexicaux spécifiques. La désambiguïsation sémantique reste néanmoins une tâche difficile, forcément limitée par les ressources linguistiques à disposition, ce qui constitue une limitation notable. Enfin, le cluste-

ring spectral utilisé pour l'inférence des types est un algorithme paramétrique dont les auteurs ne donnent pas les valeurs retenues, rendant la reproduction de leurs résultats d'autant plus difficile que le système est intrinsèquement complexe.

Les approches symboliques ne sont pas complètement abandonnées au profit de modèles probabilistes intégrés. Ainsi, AHN (2017) cherche à induire les rôles avant d'identifier les types (à l'inverse des autres propositions dans le domaine), observant que ce découpage préalable du corpus réduit le volume de données disponibles pour l'induction de rôles dans chaque type. Ce faisant, le système présenté utilise le *synset* "événement" de WordNet comme amorce pour identifier les prédicats associés aux rôles. Les arguments de ces prédicats sont ensuite groupés en combinant deux scores de similarité, le premier reposant sur leur dépendances syntaxiques et leur classe d'entité (les sujets sont distingués des objets et les sujets humains sont distingués des sujets de type organisation), d'une manière similaire à celle de CHAMBERS et JURAFSKY (2011), avec la particularité d'utiliser les hyperonymes WordNet des termes rencontrés pour affiner ce processus et lui conférer un meilleur niveau de généricité. Le second score est construit afin d'identifier les relations équivalentes (sujet du verbe "mourir" ou objet du verbe "tuer").
Le système recourt alors à une étape intermédiaire pour structurer les suivantes en dégageant les séquences narratives des documents du corpus par une heuristique séparant les paragraphes ne partageant pas d'arguments coréférents pour un rôle donné. Puis, les types d'événements sont identifiés en recherchant des déclencheurs lexicaux dans ces blocs narratifs et en groupant ceux similaires qui cooccurrent dans le même bloc. On peut alors assigner un type d'événement à une mention d'un bloc narratif en fonction des déclencheurs qu'il contient et extraire les mentions correspondant à chaque *slot* en utilisant les rôles généraux.
Il est assez clair que l'hypothèse de pouvoir tirer des rôles généraux de l'ensemble du corpus fonctionne principalement sur l'homogénéité du corpus MUC et que l'intérêt de procéder ainsi afin de ne pas restreindre la taille des corpus propres à chaque type est limité dans la perspective de l'exploitation d'un corpus de grande taille. Ces limitations sont admises par l'auteure elle-même, qui s'attache par ailleurs, comme NGUYEN et al. (2015) dans une moindre mesure, à souligner le caractère obsolète du corpus MUC pour l'évaluation de ces travaux.

Pour conclure ce panorama, YUAN et al. (2018) présentent un système d'induction de schémas d'événements en deux étapes. Dans un premier temps, les instances et les types sont inférés conjointement par un modèle bayésien non-paramétrique. Contrairement aux autres modèles de ce type utilisés pour des approches similaires, ce modèle intègre aux observables l'horodatage du document, ce qui lui permet d'intégrer une dimension temporelle au processus d'inférence et de contraindre les instances à apparaître dans des fenêtres temporelles restreintes, selon l'hypothèse que chaque instance d'événement représente un pic d'information dans le temps.

Les auteurs utilisent ensuite le découpage produit par ce modèle pour apprendre, pour chaque type d'événement, un plongement lexical de chaque entité impliquée dans les instances rattachées au type, fondé sur le graphe de cooccurrence de ces entités dans les documents. Ces plongements sont enfin utilisés pour l'induction des schémas, par application d'un clustering spectral afin de grouper en un même argument les plongements d'entités similaires.

En synthèse, ces travaux sur l'induction de structures d'événements de plus haut niveau n'ont pas bénéficié de campagnes d'évaluation ayant structuré des tâches de référence comme ont pu le faire MUC, ACE ou TDT pour les travaux présentés dans les sections précédentes.

2.2.4 La question de l'évaluation

Les approches présentées ici partagent la caractéristique d'être non-supervisées ou semi-supervisées et d'ambitionner l'exploitation de grands corpus de textes. Elles prennent souvent des formes analogues à des tâches de clustering et leur évaluation présente souvent les mêmes difficultés, constituant une problématique à part entière. On peut néanmoins distinguer dans les travaux présentés trois méthodologies récurrentes. La première, plutôt qualitative, consiste à prélever un petit échantillon aléatoire parmi les sorties du système et à les faire évaluer manuellement par un ou plusieurs annotateurs, experts ou non. Cette méthode présente des limitations évidentes liées à la taille des échantillons, qui doivent rester analysables de manière approfondie par des humains, ainsi qu'à l'expertise de ces annotateurs, qui sont le plus souvent les auteurs des travaux, même si quelques travaux ont eu la possibilité de faire faire ce travail à des journalistes, et enfin en termes de couverture, certains événements ou types d'événement étant plus faciles à identifier (les attentats sont par exemple souvent plus faciles à délimiter, aussi bien au niveau des instances que du type, qu'une affaire judiciaire ou un scandale financier).

Une deuxième approche plus quantitative compare la performance du système présenté à celles de différentes baselines, présentant par conséquent l'avantage de pouvoir recourir à des métriques standardisées comme la précision, le rappel. . . Cette méthode n'échappe cependant pas à un inconvénient partagé avec la précédente, qui est la difficulté à comparer des contributions entre elles. En effet, si les métriques sont standardisées, leur mise en œuvre est toujours adaptée au besoin des auteurs, souvent lié aux spécificités de forme des sorties à évaluer, elles-mêmes conditionnées par des problématiques applicatives propres à chaque système. Ces mêmes spécificités déterminent souvent le choix des corpus de travail et d'évaluation, accentuant encore l'hétérogénéité des contributions et donc la difficulté d'une comparaison.

La dernière méthodologie se rapproche de celle utilisée pour les tâches et systèmes supervisés et consiste à comparer le système à (au moins) un système concurrent sur un protocole commun. Cette approche se concentre ici autour des travaux de Chambers, en reprenant le protocole d'évaluation défini dans (CHAMBERS et JURAFSKY, 2011), fondé sur la mesure du recouvrement entre le contenu des schémas candidats et ceux du corpus MUC. Par conséquent les travaux qui l'utilisent ((CHEUNG et al., 2013 ; NGUYEN et al., 2015 ; SHA et al., 2016 ; AHN, 2017) ainsi que (LIU et al., 2019) dans une moindre mesure), sont les seuls à se comparer les uns aux autres sur une base commune qui en fait un standard *de facto*. Bien que présentant un intérêt certain, elle est restée très marginale et s'est cantonnée au corpus MUC, qui n'est plus adapté aux réalités applicatives actuelles et contraint de fait les auteurs à de nombreux compromis : transfert du système sur un corpus de petite taille, à la variété faible, évalué sur seulement quatre *templates* réduits à quatre *slots* chacun, aucun n'illustrant toute la complexité que peuvent revêtir certains événements. Ce protocole, bien que critiquable, reste néanmoins une proposition difficile à dépasser, ce qui explique en partie qu'il ait été si réutilisé.

On constate donc que peu de travaux se comparent frontalement les uns aux autres, en partie en raison du cloisonnement de certains domaines (entre le TAL et le Web sémantique par exemple) mais aussi en raison de la variété des cadres applicatifs évoqués plus tôt. Ce manque d'émulation se trouve renforcé par l'absence, au niveau des communautés de recherche, d'un effort d'organisation de tâches articulées autour d'un corpus unifié répondant aux réalités des besoins de modélisation et de volumétrie actuels, une absence déjà signalée par des travaux antérieurs.

2.3 Conclusion

À l'issue de cette courte présentation, nous espérons avoir donné un aperçu de la grande variété des intérêts et des objectifs qui se cristallisent autour du concept d'événement et qui explique, comme l'ont déjà observé TANNIER (2014) et SPRUGNOLI et TONELLI (2017), pourquoi aucune définition précise de l'événement n'a jamais pu être fixée. L'événement est un concept que l'on adapte à ses propres besoins et non un phénomène absolu dont il s'agirait d'affiner la modélisation. Le travail présenté ici s'inscrit pleinement dans la ligne de recherche de la dernière partie, à savoir l'induction de structures génériques d'événement. Comme nous l'avons esquissé en section 1.2, notre contribution puise dans les trois niveaux de granularité présentés : nous identifions des instances d'événements à partir de leurs descriptions textuelles, dont nous isolons les mentions les plus caractéristiques afin d'en faire émerger les types sous-jacents. Nous ré-exploitons alors les principes sous-tendant l'identification d'événements atomiques pour structurer les rôles qui décriront chacun de ces types.

Pour finir, nous tentons de nous évaluer de manière à rendre nos résultats les plus pertinents possibles, en gardant à l'esprit les difficultés et limitations inhérentes à notre domaine.

3 De la mention à l'instance

> *Par exemple, vous prenez aujourd'hui. Vous comptez sept jours. Ça vous emmène dans une semaine. Eh ben on sera exactement le même jour qu'aujourd'hui.*
>
> **— Jean-Christophe Hembert**
> Kaamelott, Livre II, Sept cent quarante-quatre

Dans ce chapitre, nous décrirons la première phase de notre méthode d'induction de schémas d'événements, à savoir la construction d'instances d'événements. Celle-ci a pour objectif de regrouper des mentions textuelles d'événements issues de contenus journalistiques en instances d'événements, sous la forme d'ensembles de documents distincts dont chacun relate une même occurrence événementielle du monde réel, par exemple l'ensemble des documents dont les mentions décrivent le séisme ayant frappé la région de Katmandou le 25 Avril 2015.
Notre solution peut être rapprochée de celles de CONRAD et BENDER (2016) et de NANNI et al. (2017), qui exploitent dans un premier temps des données issues de sources considérées comme de "meilleure qualité" pour produire des résultats qui sont ensuite utilisés comme base et étendus à des corpus plus variés. Toutefois, contrairement à CONRAD et BENDER (2016), nous ne disposons pas de méta-données sur les instances que nous manipulons et que nous pourrions exploiter en supervision distante. Nous nous distinguons également du travail présenté dans (NANNI et al., 2017) car ce dernier utilise cette extension comme stratégie de repli dans le cas où la source primaire (de meilleure qualité) ne produit pas de réponse, tandis que nous l'envisageons comme une composante à part entière de notre processus de construction des instances.

Cette phase se décompose en trois étapes :

1. le clustering d'un corpus de dépêches AFP en instances d'événements ;
2. l'augmentation du volume et de la variété des descriptions liées à ces clusters par la récupération d'articles issus du Web et leur agrégation autour des clusters AFP en utilisant ces derniers comme amorces ;
3. la consolidation des clusters obtenus par le filtrage des articles Web les moins similaires aux amorces AFP.

Les sections suivantes présentent d'abord les données que nous utilisons, puis détaillent chacune des étapes mentionnées ci-dessus, dans l'ordre de leur énumération.

3.1 Présentation des corpus de travail

Corpus AFP

Comme nous l'avons mentionné en section 1.4.2, nous avons eu accès pour nos travaux à un corpus de dépêches AFP, ressource textuelle qui est le produit d'un processus de rédaction particulier. En effet, en raison du caractère imprévisible de l'occurrence de la plupart des événements et de la nécessité de les relayer rapidement, une instance d'événement est souvent décrite à travers un fil de dépêches. La première, appelée *alerte*, se réduit souvent à un titre seul, puis une deuxième dépêche est émise à sa suite, comportant un chapeau, parfois un corps succinct. Enfin, le corps est développé dans les dépêches suivantes, ajoutant ou corrigeant des informations, comme par exemple les bilans de victimes. Une illustration de ce processus est disponible en figure 1.4 de la section 1.4.2.

Si ce processus tend indéniablement à produire des mentions redondantes, surtout au niveau des titres et des chapeaux, qui se fixent assez rapidement et se trouvent donc répercutés à l'identique dans la suite du fil, une variabilité plus ou moins importante reste néanmoins présente et à exploiter. *A contrario*, beaucoup d'événements, par leur caractère exceptionnel ou important, se voient couverts de différentes façons, soit par des auteurs multiples, soit sous des angles différents. Cela se traduit par une variété significative dans les descriptions, comme on peut le voir sur la figure 3.1.

Corpus Web

Au cours de cette première phase, nous exploiterons également un corpus d'articles issus du Web et déjà introduit en section 1.4.5. Le tableau visible en annexe B détaille les sources utilisées.

3.2 Clustering en instances du corpus AFP

Dans cette première étape, nous cherchons à obtenir des descriptions d'instances "pures" en exploitant les spécificités des dépêches AFP, à savoir la rigueur des règles

Fig. 3.1: Illustration de la variété des mentions liées à une même instance. Ici, à quelques minutes d'intervalle, deux auteurs émettent deux dépêches présentant l'événement sous deux angles différents (avec néanmoins un élément commun, l'absence de survivants).

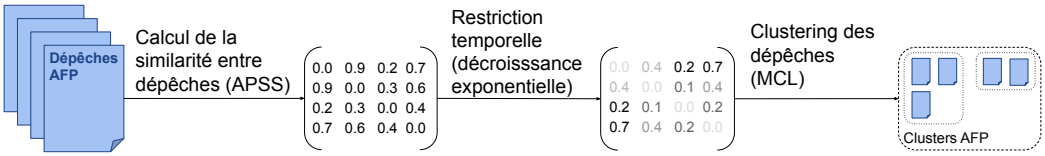


Fig. 3.2: Synthèse des différentes opérations conduisant à la construction d'instances AFP, clusters de dépêches décrivant la même instance d'événement.

qui régissent leur rédaction et le processus itératif de leur production, qui induisent une forte similarité entre deux dépêches décrivant la même instance, ainsi que la couverture multiple de certains événements, qui permet de capturer une certaine variabilité lexicale pour une description d'instance donnée. Une synthèse de ce processus est donnée par la figure 3.2.

Pour ce faire, nous adoptons une approche de Recherche d'Information (RI) classique fondée sur des techniques de clustering de documents. Nous utilisons un algorithme non-paramétrique (c'est-à-dire ne requérant pas la spécification d'un nombre de cluster fixe donné à l'avance) reposant sur la similarité entre documents. Afin de mettre en œuvre cette technique, nous commençons par construire une représentation parcimonieuse binaire de chaque document du corpus. Nous retenons cette représentation vectorielle simple (par rapport à l'application d'un schéma de pondération de type tf.idf) pour deux raisons.

Nous pensons d'abord qu'elle suffit à représenter nos documents compte tenu de leur petite taille individuelle. En effet, se réduisant essentiellement à deux phrases, il y a peu de répétition entre les termes d'un document, ce qui réduit l'amplitude des effets d'un schéma de pondération et donc son utilité, sans pour autant réduire son coût de calcul. De plus, nous savons que l'identification d'instances au sein du

corpus AFP implique de travailler sur des documents très similaires. Le choix d'une représentation binaire permet de définir des frontières plus nettes entre documents identiques ou quasi-identiques, vraisemblablement traitant de la même instance, et des documents thématiquement similaires mais traitant d'une instance différente. Chaque dépêche se trouve ainsi représentée par un vecteur à $|V|$ dimensions, où $|V|$ est la taille du vocabulaire V du corpus, dans lequel la dimension t sera positionnée à 1 si le terme t est présent dans la dépêche, 0 sinon, $\forall t \in V$. Nous restreignons la constitution de ce vocabulaire aux lemmes des noms (propres et communs), adjectifs et verbes, termes par nature les plus à même de décrire un événement. La lemmatisation est utilisée pour favoriser l'identification de termes identiques.

Nous calculons ensuite la matrice de similarité cosinus des documents ainsi représentés à l'aide de l'algorithme All Pair Similarity Search (APSS) (BAYARDO et al., 2007). La similarité cosinus a été choisie car il s'agit d'une mesure simple, bien établie et correspondant à nos besoins pour cette étape, à savoir comparer des documents potentiellement redondants selon une approche sac de mots. L'algorithme APSS a lui été retenu car il est adapté au traitement de représentations parcimonieuses binaires et possède une grande capacité de passage à l'échelle, une caractéristique à prendre en compte étant donné les volumes que nous traitons. Il a par ailleurs l'avantage de calculer une valeur exacte de la similarité cosinus, contrairement à d'autres algorithmes de recherche efficace des plus proches voisins qui se fondent sur un calcul approché pouvant poser problème. APSS est régi par un seul paramètre fourni par l'utilisateur, à savoir le seuil de similarité minimale en dessous duquel une paire d'éléments, ici des dépêches, n'est pas conservée, ce seuil influençant la performance de l'algorithme (plus le seuil est bas, plus l'exécution sera longue).

La caractéristique principale d'une instance d'événement, en comparaison d'une mention ou d'un type, est son ancrage dans un contexte spatial et temporel. La majorité des travaux partageant cette vision de l'événement l'utilise de façon explicite, que ce soit pour fixer une limite à l'existence d'un événement, comme (AZZOPARDI et STAFF, 2012) ou (CONRAD et BENDER, 2016) ou pour reconstituer un enchaînement causal des sous-événements qui composent l'instance, à l'instar de (ROSPOCHER et al., 2016) ou (GLAVAŠ et ŠNAJDER, 2013). Nous tenons compte de cette caractéristique temporelle en pondérant le score de similarité cosinus obtenu entre chaque paire de dépêches par l'application d'une décroissance exponentielle fondée sur le décalage temporel entre les deux documents considérés. Ainsi, à similarité cosinus égale, deux documents éloignés dans le temps verront leur similarité finale diminuer comparativement à deux documents chronologiquement plus proches. Le score de similarité final entre deux documents d_1 et d_2 prend donc la forme :

$$sim(d_1, d_2) = \frac{cos(d_1, d_2)}{e^{\delta/\omega}}$$

où $cos(d_1, d_2)$ désigne la similarité cosinus “brute" entre les deux documents considérés, δ est la valeur absolue de la différence entre les dates de création des deux documents, en heures, et ω définit la largeur de la fenêtre temporelle.

Cette pondération vise à prendre en compte une forme “d'inertie" médiatique de l'événement, à savoir que certaines instances d'événements peuvent être la source d'un flux continu d'informations connexes pendant plusieurs heures, voire plusieurs jours. Un exemple simple de ce phénomène se trouve dans les événements de type catastrophe tels que les séismes : le séisme est une instance bien définie, qui entraîne des victimes, dont le bilan va évoluer pendant un certain temps avant de se stabiliser. Chaque mise à jour de ce bilan est associée à l'émission d'une nouvelle dépêche. La durée de ce processus dépend du type de l'événement (lors d'un accident d'avion, le bilan se stabilise généralement plus vite que pour un séisme par exemple) et de l'instance considérée (un séisme particulier peut présenter des circonstances à même de retarder l'action des secours). C'est ce flux, cette inertie, que nous souhaitons mieux capturer par la définition de cette fenêtre temporelle. Car, indépendamment de la question de savoir s'il faut ou non grouper les documents relatant le séisme avec les bilans et leur mise à jour (ou plus généralement la mention d'une instance et un flux informationnel dont elle est la cause), il est évident qu'il est souhaitable de grouper toutes les dépêches relayant les bilans, quand bien même leur émission se répartit sur plusieurs jours.

Enfin, nous formons des clusters de dépêches en appliquant l'algorithme Markov Clustering (MCL) (DONGEN, 2000) sur la matrice de similarité après application de la pondération temporelle. Le MCL procède au partitionnement d'un graphe en y simulant des écoulements de flux par une alternance de marches aléatoires et de phases d'*inflation* consistant à renforcer les probabilités de transitions fortes et à affaiblir les probabilités de transitions faibles. Cet algorithme a été choisi car il est adapté au traitement d'une matrice de similarité et ne nécessite pas qu'un nombre de classes lui soit spécifié *a priori*, contrairement à l'algorithme K-means par exemple. Cette caractéristique est importante car étant donnée la forte variabilité du nombre, de la nature et de l'intensité des événements relayés par l'AFP d'une période temporelle à une autre (et ce, aussi bien à l'échelle du jour que du mois ou autre), la définition d'un nombre d'instances fixe serait un paramètre à la fois difficile à optimiser et limitant fortement la qualité des résultats finaux.

Pour résumer, cette première étape exploite donc les spécificités des dépêches AFP ainsi que les informations temporelles qui les décrivent pour produire des clusters

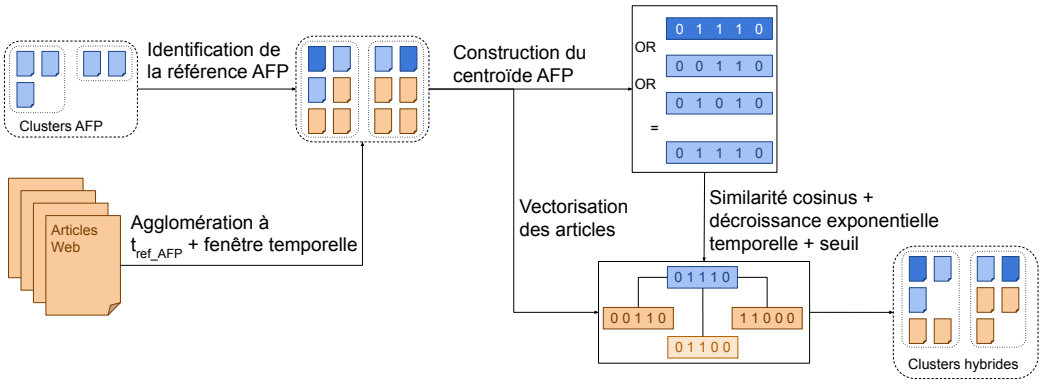


Fig. 3.3: Schéma décrivant les différentes opérations pour l'agrégation d'articles Web autour des clusters AFP.

regroupant plusieurs dépêches décrivant une même instance d'événement. Dans la suite de ce document, nous désignerons ces clusters de dépêches sous le nom de *clusters AFP*.

3.3 Agrégation d'articles issus du Web autour d'amorces AFP

Dans cette étape, nous utilisons les clusters AFP construits à l'étape précédente comme amorces pour la récupération de mentions issues d'articles Web. Partant du constat que les dépêches AFP sont d'une grande qualité rédactionnelle mais redondantes sur le plan lexical, en raison des contraintes sur la rédaction des chapeaux et du processus de production itératif dont elles sont issues, l'objectif est ici d'incorporer la variété lexicale manquante par le biais d'articles de presse généraliste en jouant à la fois sur la diversité des sources et sur le fait que, ne s'agissant pas de dépêches d'agences, les contraintes rédactionnelles y sont moins fortes. Un schéma global de cette étape est représenté en figure 3.3. Cette approche consistant à étendre à partir d'un corpus de grande taille et moins focalisé des représentations plus précises mais moins nombreuses peut être rapprochée de celle de CHAMBERS et JURAFSKY (2011), bien que cette stratégie y soit utilisée pour l'induction directe de schémas et non d'instances.

Pour ce faire, nous construisons la représentation d'un cluster AFP comme le OU logique de la représentation parcimonieuse binaire de chacune des dépêches qui le compose. Nous nommons cette représentation le *centroïde* du cluster AFP.

L'agrégation se fait dans un premier temps sur la base des horodatages des dépêches et des articles : pour chaque cluster AFP, on identifie la dépêche la plus ancienne

chronologiquement, que l'on nomme *référence AFP*, et on récupère tous les documents Web se trouvant dans une fenêtre temporelle égale à celle utilisée lors de la pondération de la similarité cosinus par décroissance exponentielle de l'étape précédente.

Bien entendu, ce premier regroupement purement temporel requiert un filtrage complémentaire sur le contenu des documents. Une double difficulté se pose alors : compte tenu des différences stylistiques significatives entre les dépêches d'agences et les articles de presse généraliste (ces différences sont discutées plus en détail à la section 3.5.2), ainsi que de la plus grande variété du vocabulaire utilisé dans les documents Web, il est nécessaire de recourir à une méthode de comparaison autorisant plus de souplesse qu'une similarité cosinus simple se fondant sur des identités de formes. De plus, une telle méthode sera plus coûteuse et ne pourra être appliquée à l'ensemble des documents agrégés autour de chaque centroïde, d'autant que beaucoup de ces documents n'ont probablement aucun rapport avec l'instance décrite dans le centroïde, n'ayant été agrégés qu'en raison de leur correspondance temporelle.

Pour pallier cette difficulté, nous souhaitons opérer un premier filtrage grossier fondé sur la similarité cosinus afin de réduire le nombre de documents sur lesquels s'appliquera notre similarité "souple". On construit donc la représentation parcimonieuse binaire de chaque article Web, pour laquelle on calcule la similarité cosinus avec le centroïde. La décroissance temporelle est ensuite appliquée, de la même façon qu'à l'étape précédente. Considérant que nous comparons la représentation d'un document à celle de plusieurs documents combinés et que nous souhaitons essentiellement évacuer la masse des documents non pertinents, le choix d'un seuil de similarité très bas s'impose afin de préserver un rappel élevé.

Dans cette deuxième étape, nous enrichissons donc les descriptions d'instances fournies par nos clusters AFP à l'aide de mentions issues d'articles Web afin de compenser la redondance inhérente aux dépêches AFP. Ce faisant nous avons introduit une certaine quantité de bruit dans nos descriptions que nous allons éliminer en deux temps, le premier consistant en un filtrage fondé sur la similarité cosinus avec un seuil très bas et présenté ici, le second par le calcul d'une similarité tenant compte de la variété lexicale capturée, détaillée dans la partie suivante. Dans la suite de ce manuscrit, nous appelons le résultat de cette étape des *clusters hybrides*.

3.4 Filtrage du contenu Web récupéré

L'objectif de cette étape finale est de filtrer les articles Web des clusters hybrides obtenus précédemment pour ne conserver que les plus similaires aux dépêches AFP servant d'amorces. La figure 3.4 présente l'ensemble des opérations effectuées. Ce filtrage requiert une conception souple de la similarité entre documents, en raison des différences stylistiques importantes pouvant exister entre une dépêche AFP et un article de presse généraliste, même en se cantonnant au titre et à la première phrase de chaque document. L'étude de la similarité entre une dépêche et un article relève dès lors moins de la quantification de termes identiques que de la mesure du degré de paraphrase entre l'une et l'autre, selon une dimension sémantique plus large. En conséquence, nous abandonnons la similarité cosinus pour adopter une mesure de similarité tenant compte de cet aspect paraphrastique. Nous utilisons à cette fin le système d'alignement monolingue de SULTAN et al. (2014). Ce système aligne deux phrases mot à mot par l'application de plusieurs modules dédiés chacun à un type d'unité linguistique particulier. Nous l'avons choisi pour plusieurs raisons : en premier lieu, il s'agit du système ayant obtenu les meilleures performances lors de la campagne SemEval 2014. Il est par ailleurs léger à déployer et facile à modifier, ce qui nous permet de l'adapter à nos besoins. Enfin, ce système produit permet de récupérer les paires de termes alignés, ce qui constitue une sortie plus riche pour nos objectifs qu'un simple score de similarité entre deux phrases.

Le système commence par aligner tous les n-grammes ($n \geqslant 2$) de mots contenant au moins un mot plein (selon une liste de mots vides prédéfinie). L'alignement de deux n-grammes est fait si leurs chaînes de caractères sont identiques. Puis les entités nommées sont considérées et alignées en cas de correspondance partielle des chaînes de caractères, le système étant capable d'aligner un acronyme avec son expansion par exemple (en utilisant l'outil *Stanford NER* (FINKEL et al., 2005)).

Les mots pleins non encore alignés sont ensuite traités par le calcul d'un score de similarité fondé sur la recherche de deux types d'indices contextuels : syntaxiques et lexicaux. Pour chaque paire de mots pleins candidate à l'alignement, le système recherche des appariements sur la base d'équivalences entre relations de dépendance syntaxique liant les paires (dans leurs phrases respectives). Ces équivalences ont été établies à la main et sont exploitées dans le système sous forme de règles, l'analyse syntaxique en dépendances est intégralement réalisée avec le *Stanford Parser* (MARNEFFE et al., 2006 ; TOUTANOVA et al., 2003). Les indices lexicaux sont quant à eux recherchés dans une fenêtre de 6 termes (3 précédents et 3 suivants) entourant le terme candidat, sous la forme d'équivalences sémantiques. Ces équivalences sont trouvées si deux termes de cette fenêtre sont identiques dans chaque phrase ou

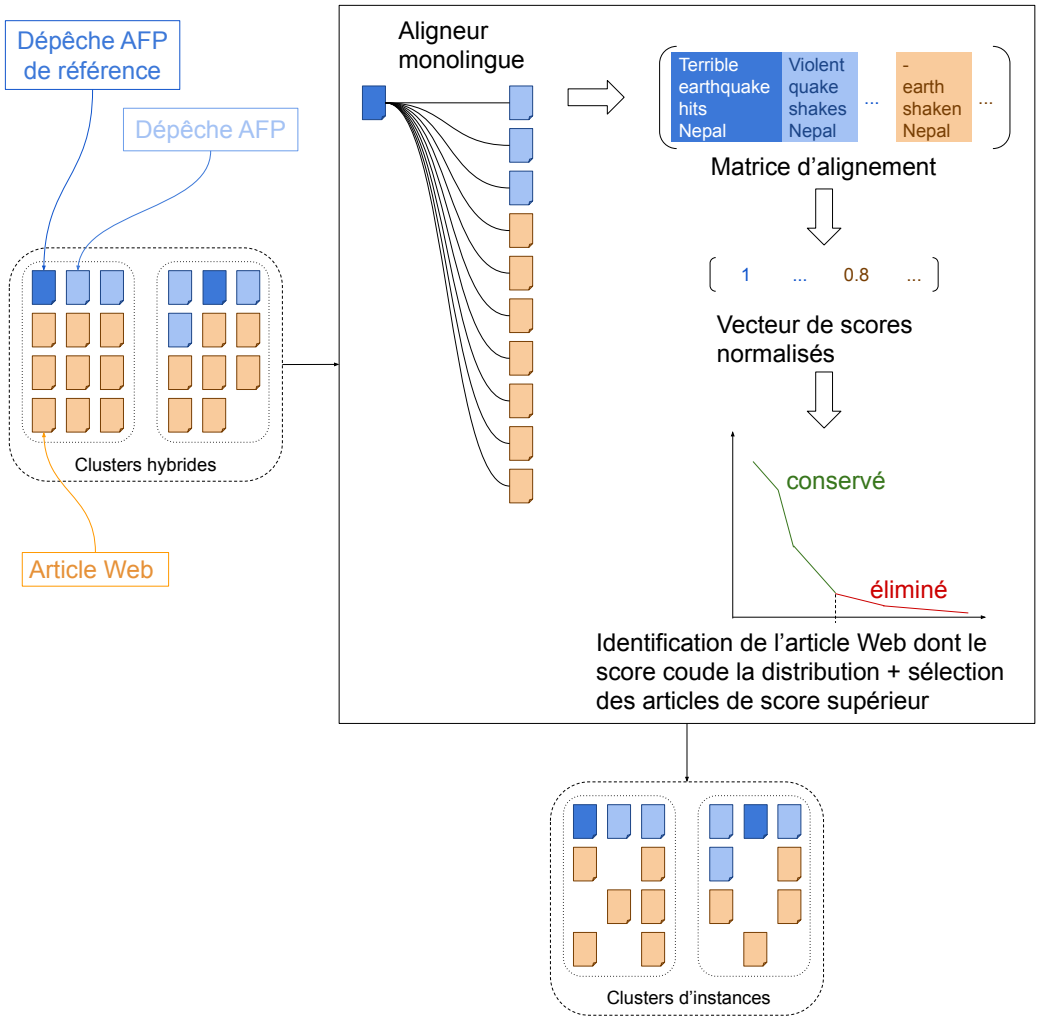


Fig. 3.4: Synthèse du processus de filtrage des clusters hybrides.

s'ils sont appairés dans la base Paraphrase Database (PPDB) [1] (GANITKEVITCH et al., 2013), une ressource qui contient notamment de nombreuses paires de termes synonymes acquises à partir de corpus parallèles. Un score combinant ces deux types d'indices est ensuite calculé et chaque terme se voit aligné avec celui ayant le meilleur score. Enfin, le même traitement est appliqué aux mots vides non encore alignés.

Nous avons par ailleurs adapté ce système afin qu'il corresponde mieux à nos besoins en redéfinissant la liste des mots vides afin notamment d'y intégrer les jours de la semaine, très utilisés dans notre corpus. Nous avons également étendu la recherche de n-grammes identiques aux unigrammes et utilisé l'absence d'alignements à l'issue de ce premier module et du suivant (dédié aux entités nommées) comme condition d'arrêt précoce du processus d'alignement. En effet, bien que nous souhaitions exploiter une forme de similarité souple, nous n'en traitons pas moins des instances

1. Disponible en téléchargement libre à l'adresse `http://paraphrase.org/#/download`

d'événements, dans lesquelles certains éléments (par exemple les mentions géographiques) doivent nécessairement se recouper. Dans cette optique, il ne paraît pas raisonnable de chercher des correspondances distantes (et coûteuses) entre deux documents ne partageant aucun terme en commun. Dans la continuité de ce raisonnement, le processus d'alignement n'est appliqué que si les deux documents considérés partagent une entité nommée de type `LOCATION`. Enfin, les termes manipulés ne sont pas les formes de surface issues des documents sous-jacents mais une combinaison de leurs lemmes et de leurs étiquettes en partie du discours.

Ce système est utilisé pour produire les alignements entre la référence AFP (la dépêche la plus ancienne dans l'ordre chronologique) du cluster AFP et tous les autres documents du cluster hybride, c'est-à-dire à la fois les autres dépêches du cluster AFP et les documents Web agrégés et ce, pour chaque cluster hybride. Rappelons que chaque document se compose en réalité de deux phrases : le titre et le chapeau. Nous commençons par aligner le titre de la référence avec celui du document candidat puis procédons de même avec les chapeaux. L'opération d'alignement étant assez coûteuse en termes de temps, nous privilégions la comparaison de chaque document à une même "ancre" pour éviter une explosion du temps de calcul. De plus, l'alignement reposant en partie sur des critères syntaxiques, il est nécessaire que cette ancre soit un document réel plutôt que le centroïde, qui est un sac de mots. Nous attribuons ce rôle à la référence AFP car, étant la première chronologiquement, elle ne contient que l'essentiel de l'information décrivant l'événement considéré. Il semble donc raisonnable de considérer que tous les autres documents auront un lien paraphrastique plus fort avec celle-ci qu'avec n'importe quel autre. Pour des raisons de passage à l'échelle, nous n'appliquons ce traitement qu'aux 50 premiers articles de chaque cluster après avoir constaté que si les articles étaient traités par ordre décroissant de similarité cosinus, il était extrêmement rare de trouver plus de 50 articles d'intérêt.

Ces alignements successifs nous permettent de construire une matrice d'alignement de termes dont chaque colonne est associée à un document et chaque ligne représente l'alignement d'un terme de la référence AFP dans tous les autres documents.

Un score est ensuite calculé pour chaque document en tenant compte du nombre de termes de ce document ayant trouvé un alignement avec la référence AFP ainsi que du nombre de documents dans lequel le terme de référence a été aligné. Pour cela, nous commençons par binariser cette matrice : si un terme t du document d a été aligné avec un terme de la référence AFP, alors $B_{t,d} = 1$, sinon $B_{t,d} = 0$.

Nous construisons alors le vecteur de poids $Align$ associé aux termes alignés, de la forme :

$$Align_d = \sum_{t=0}^{N} B_{t,d}$$

où N est le nombre de termes de référence ayant trouvé au moins un alignement.

Le score d'alignement absolu de chaque document d est donné par :

$$s_d = \sum_{t=0}^{M} W_{t,d}$$

où M est le nombre de documents alignés $W = B \times Align^T$.

Ce score est normalisé par celui qu'aurait obtenu la référence AFP. Pour cela, nous construisons le vecteur de poids $Norm$ en reprenant $Align$ et en y remplaçant les valeurs nulles par des 1. La figure 3.5 synthétise graphiquement le processus de calcul du score d'alignement.

Enfin, l'ensemble d'articles Web le plus pertinent par rapport à l'amorce AFP est sélectionné par le calcul du "coude" de la distribution des scores d'alignement. Ce critère a été choisi car il permet, par un calcul simple, de déterminer un seuil adapté à la distribution des scores de chaque cluster, laquelle peut fortement varier d'un cluster à un autre. Cette valeur est déterminée par le plus grand écart entre deux scores d'alignement consécutifs sur l'ensemble des scores d'alignement triés par ordre décroissant. Nous conservons les articles dont le score est supérieur à ce seuil. Il faut préciser qu'il peut arriver que certaines sources Web reprennent une dépêche AFP en introduction de leur article, notamment dans les chapeaux, ce qui se traduira par un score d'alignement anormalement élevé, faussant la distribution des scores et donc le calcul du coude, qui ne conservera alors que ces "reprises". Pour pallier ce problème, nous opérons le calcul du coude en excluant les articles dont le score est supérieur à 0,90, seuil empirique dont nous avons constaté qu'il n'était atteint que dans ce cas particulier.

Pour résumer, nous utilisons dans cette étape un système d'alignement monolingue pour établir une forme souple de similarité entre la dépêche AFP la plus ancienne d'un cluster hybride et tous les autres documents de celui-ci. Nous exploitons ces alignements pour constituer une matrice grâce à laquelle nous établissons pour chaque document un score d'alignement tenant compte du nombre de termes alignés dans celui-ci et du nombre de fois où ce même terme a trouvé un alignement à travers l'ensemble des documents. Nous identifions le coude dans la distribution

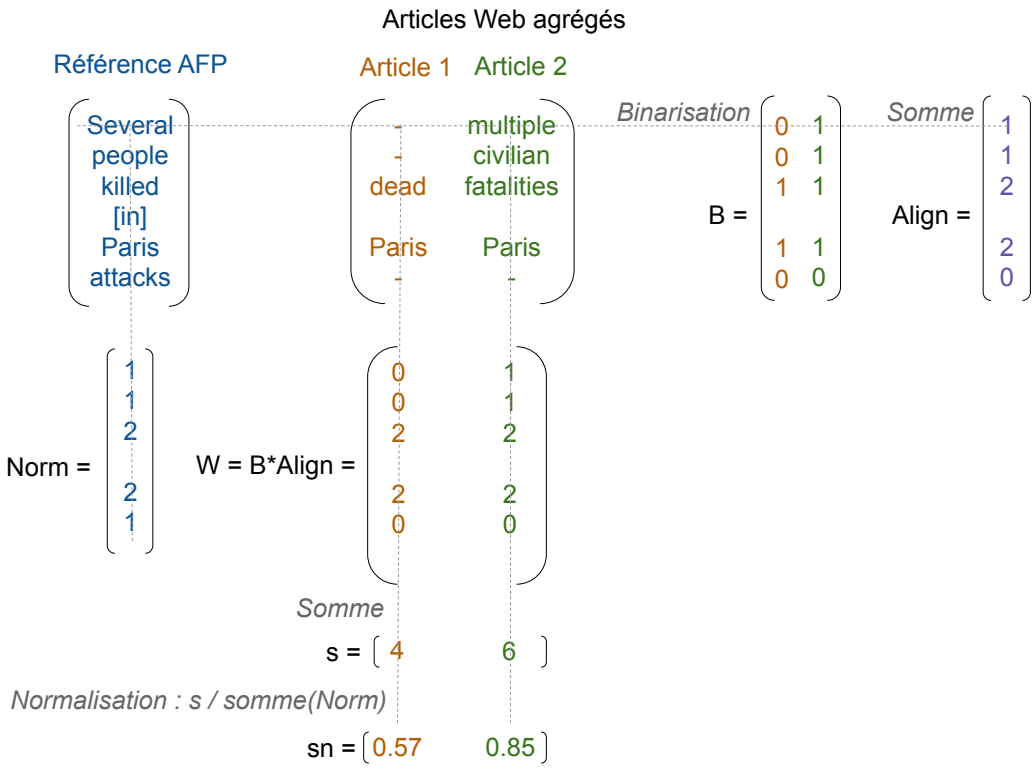


Fig. 3.5: Représentation de l'opération de calcul du score d'alignement de chaque document avec la référence AFP.

des scores résultants pour exclure les documents non pertinents. En répétant cette opération pour chaque cluster hybride, nous obtenons des ensembles de mentions mieux focalisés sur la même instance d'événement. Dans la suite de ce document, nous ferons référence à ces ensembles sous le terme de *descriptions d'instances*.

3.5 Évaluation

3.5.1 Clustering en instances du corpus AFP

Méthodologie

Le regroupement des dépêches reposant essentiellement sur leur similarité cosinus, l'influence du seuil de similarité fixé pour l'APSS est un paramètre primordial à explorer. Il conditionne la forme de la matrice de similarité résultante (plus ou moins creuse) sur laquelle s'applique le MCL.

Cette similarité est ensuite pondérée par une décroissance exponentielle dont la largeur de la fenêtre temporelle détermine principalement les valeurs finales de la matrice de similarité, tout en contribuant possiblement à la rendre plus creuse.

Concernant le MCL, bien que cet algorithme ne nécessite pas qu'on lui spécifie *a priori* un nombre de classes à constituer, il dispose d'un paramètre appelé facteur d'inflation, influant sur la taille des classes obtenues. Plus le facteur d'inflation est grand, plus les classes formées auront tendance à être petites, ne regroupant que peu de documents, toutes choses égales par ailleurs.

Ces trois paramètres (seuil de similarité, fenêtre temporelle de la décroissance exponentielle et facteur d'inflation) ayant tous une influence directe sur le comportement de l'algorithme et les clusters en résultant, nous avons décidé de les optimiser conjointement par recherche exhaustive, autrement dit par une recherche par grille (*grid search*).

Nous explorons tous les seuils de similarité entre 0,1 et 0,8 par pas de 0,1. Les valeurs supérieures à 0,8 correspondent en pratique à un regroupement des doublons uniquement, ce qui ne nous intéresse pas. C'est pourquoi elles ont été écartées.

Nous combinons ces valeurs de seuil à six largeurs de fenêtres temporelles : 0 (pas de décroissance appliquée, la similarité cosinus est exploitée telle quelle), 1 jour, 2 jours, 3 jours, 7 jours et 15 jours. Nous avons d'abord choisi des valeurs proches les unes des autres (entre 1 et 3 jours) pour capturer une inertie "locale", par exemple l'évolution des bilans d'une attaque terroriste ou d'un accident d'avion mais aussi des valeurs plus lointaines (7 et 15 jours) pour des événements qui produisent des flux d'informations plus sporadiques mais sur une durée plus longue, par exemple les bilans provisoires d'un séisme, en particulier dans le cas où l'action des secours est rendue difficile. Dans ce cas, le flux global de production de dépêches reprend une activité "normale" pendant plusieurs jours, c'est-à-dire relatant une variété d'événements différents, puis ré-émet une nouvelle série de dépêches centrées sur le bilan relatif au séisme.

Pour finir, l'influence du facteur d'inflation étant jugée de plus faible importance et l'ensemble des valeurs pertinentes à explorer étant plus difficile à définir, nous nous cantonnons aux valeurs recommandées par l'auteur [2], à savoir 1,4, 2, 4 et 6.

Nous évaluons ensuite la qualité des instances identifiées par notre système en comparant les dépêches de chaque cluster avec celles étiquetées dans le cadre des travaux de RUDNIK et al. (2019), qui mettent en correspondance des dépêches

2. À l'adresse `https://micans.org/mcl/`.

AFP avec des instances d'événements dans Wikidata (*item* dans la nomenclature Wikidata).

Ce processus de mise en correspondance est effectué automatiquement et repose sur le calcul d'un score de similarité fondé sur la recherche, dans le texte des dépêches AFP, de mentions utilisées pour remplir certains champs décrivant l'événement dans Wikidata, à savoir :

— les champs liés à la temporalité de l'événement : `point_in_time` si l'événement est ponctuel ou `start_time` et `end_time` s'il s'inscrit dans une durée ;

— les champs liés à la localisation de l'événement, c'est-à-dire les champs `country` et `location` ;

— les champs liés au sujet de l'événement. Pour ce faire, le titre de l'*item* est exploité et complété par le *Wikidata Event Type* (WET). Le WET se définit par les valeurs du champ `instance_of` de l'*item*. Ces éléments sont utilisés comme liste de mots-clés. Un exemple d'*item* Wikidata et de WETs associés est visible en figure 3.6.

Cette ressource est particulièrement intéressante pour notre objectif en raison de sa grande qualité. En effet, les auteurs rapportent une précision de 100 % en opérant la recherche des termes identifiés dans Wikidata sur les 3 premières phrases de chaque dépêche au prix d'un rappel faible (67 %), ce qui nous semble un compromis acceptable.

La tâche d'évaluation de résultats de clustering est difficile, comme en témoigne la multiplicité des mesures disponibles, qu'elles soient internes, c'est-à-dire qu'elles évaluent la qualité d'un clustering candidat par l'étude de la cohérence de ses clusters, ou externes, c'est-à-dire qu'elles comparent le clustering candidat à une référence. Nous privilégions une mesure externe car les mesures internes (Silhouette, Dunn...) se fondent sur la quantification du contraste entre les distances intra- et inter- classes. Or, cette approche n'est pas adaptée à la comparaison de nos différentes configurations, dont chaque paramètre agit comme un filtre susceptible d'éliminer un nombre plus ou moins important de documents du clustering final. Le paramètre illustrant le plus clairement ce phénomène est le seuil de similarité de l'APSS. Toute paire de documents dont la similarité est inférieure au seuil est éliminée et nous faisons varier ce dernier entre 0,1 et 0,8. Il est donc certain que le nombre de documents filtrés pour la valeur 0,8 sera beaucoup plus important que pour la valeur 0,1, ce qui constitue un biais en faveur des configurations les plus restrictives dans le cas du recours à une mesure interne.

Nous fondons donc notre évaluation sur le calcul de la pureté, un critère d'évaluation externe mesurant la correspondance entre deux clusters, l'un servant de référence, l'autre étant le candidat évalué. Plus précisément, nous calculons la micro- et macro-

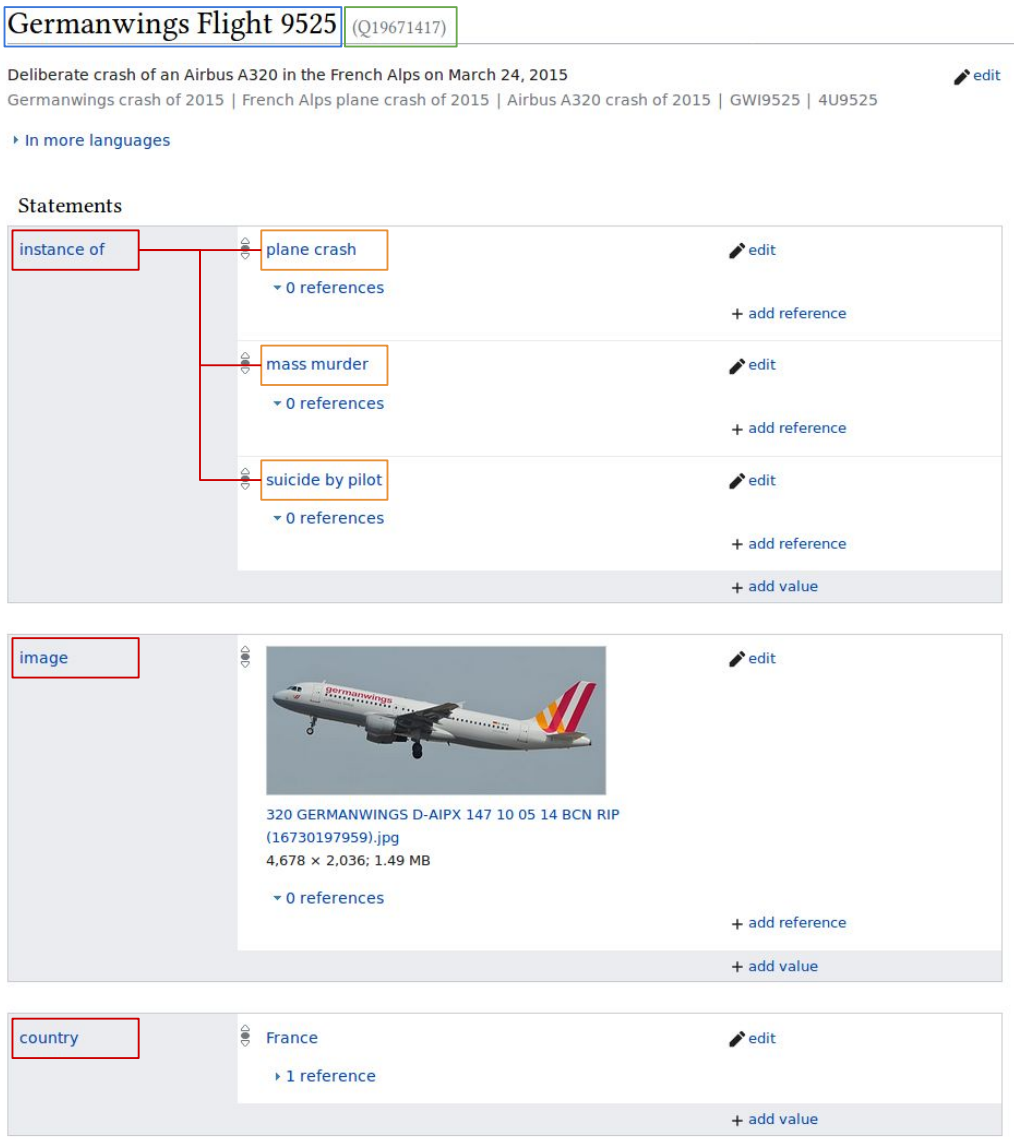


Fig. 3.6: Capture de l'*item* Wikidata associé au crash du vol 9525 de la Germanwings. Dans le cadre vert en haut, l'identifiant de l'*item*, en bleu, le titre de l'*item* ; en rouge, les différents champs qui le décrivent. Le *Wikidata Event Type* (WET) est le contenu du champ *instance_of*. Ici, cet accident se rattache à trois WETs : *plane crash*, *mass murder* et *suicide by pilot*, en orange.

puretés ainsi que la micro- et macro-puretés inverses entre les deux ensembles de clusters formés d'une part, par notre système et d'autre part, par la correspondance de référence AFP/Wikidata dans laquelle chaque cluster est formé par les dépêches associées à un même *item*. Nous choisissons la pureté parmi d'autres mesures externes car il s'agit d'une mesure simple, bien établie et facilement interprétable. L'exploitation de la pureté inverse permet de plus de maîtriser son biais pour les clustering fins. En effet, la pureté se fonde uniquement sur le nombre de représentants de la classe de référence majoritaire dans le cluster candidat. Ainsi, un clustering

répartissant les éléments d'un cluster de référence entre plusieurs clusters candidats obtiendra un score élevé malgré sa granularité plus fine. La pureté inverse permet un meilleur regard sur ce comportement et privilégier, si besoin, une configuration mieux équilibrée, à pureté comparable. Pour un cluster candidat k et un cluster de référence r, la pureté se définit comme suit :

$$\textit{pureté}(k) = \frac{|l_m|}{|k|}$$

où $|l_m|$ = nombre d'éléments du cluster de référence r majoritairement représentés dans le cluster candidat k et $|k|$ = nombre d'éléments dans le cluster candidat k.

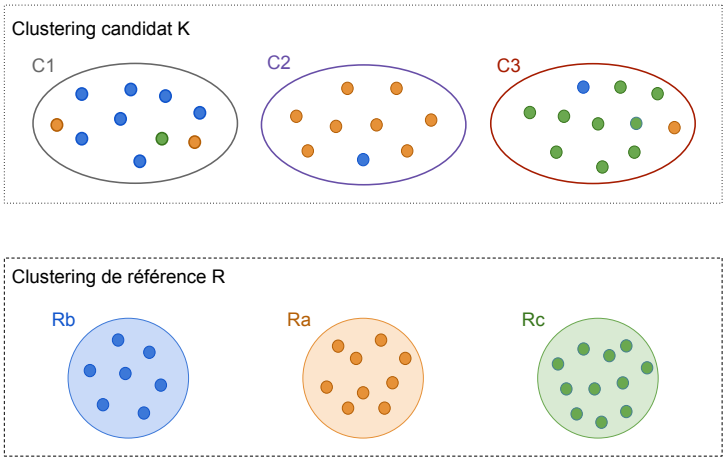
Appliquée à un ensemble de K clusters (par exemple le résultat d'un algorithme de clustering), on peut en exprimer deux variantes :

$$\textit{macro pureté}(K) = \frac{1}{K} \times \sum_{k=1}^{K} \textit{pureté(k)}$$

$$\textit{micro pureté}(K) = \sum_{k=1}^{K} \textit{pureté(k)} \times \frac{|k|}{\sum_{k=1}^{K} |k|}$$

La pureté inverse consiste simplement à inverser la référence et le candidat. Une représentation schématique de la pureté ainsi qu'un exemple sont visibles en figure 3.7.

Nous appliquons dans un premier temps ce protocole d'évaluation à l'ensemble des dépêches émises au cours de l'année 2015 et associées à au moins un des Subject Codes IPTC nous servant de filtres (cf. section 1.4.4), puis utilisons ces résultats pour sélectionner une configuration de paramètres que nous appliquerons à l'ensemble des dépêches émises sur les années 2013, 2014 et 2016 (filtrées selon les mêmes critères). Les résultats obtenus seront évalués à leur tour. L'année 2015 est exploitée de façon isolée car elle contient un grand nombre de dépêches, à la fois pour nos travaux et dans l'ensemble des données de référence associées à Wikidata, les trois autres années formant l'autre grand ensemble de données à notre disposition. Le détail de leurs volumétries respectives est donné dans le tableau 3.1. Cette approche consistant à utiliser les données de l'année 2015 comme ensemble de validation et des années 2013, 2014 et 2016 comme ensemble de test sera par ailleurs réitérée à chaque étape de notre travail.



$$pureté(C1) = \frac{7}{10} = 0.70$$

$$macro\ pureté(K) = \frac{1}{N} \times \sum_{n=1}^{N} pu(Cn) = \frac{\frac{7}{10}+\frac{8}{9}+\frac{9}{11}}{3} = 0.80$$

$$micro\ pureté(K) = \sum_{k=0}^{N} pureté(k) \times \frac{|k|}{\sum_{k=0}^{K}|k|} = 0.70 \times \frac{10}{31} + 0.88 \times \frac{9}{31} + 0.83 \times \frac{12}{31} = 0.80$$

Fig. 3.7: Illustration de la pureté comme mesure d'évaluation. Soit un clustering K constitué de $N = 3$ clusters $C1$, $C2$ et $C3$, comparé à un clustering de référence R composé lui aussi de trois clusters : Ra, Rb et Rc. Pour chaque cluster candidat Cn, la pureté se définit comme le quotient du nombre d'occurrences de la classe de référence majoritaire dans le cluster candidat par le nombre total d'éléments dudit cluster. La macro-pureté agrège ce résultat en sommant les puretés de chaque cluster et en normalisant. La micro-pureté pondère cette valeur par la proportion de documents que contient chaque cluster.

	2015	2013	2014	2016
Nb dépêches disponibles pour nos travaux	52 724	39 974	38 331	49 436
Nb dépêches disponibles pour l'évaluation des clusters AFP	38 457	4 115	5 844	5 895

Tab. 3.1: Détail de la volumétrie des données AFP disponibles sur les années 2013 à 2016. L'année 2015 est donnée en premier car elle constitue notre ensemble de validation, les trois autres constituant notre ensemble de test.

Résultats et discussion

La recherche exhaustive décrite précédemment combine 8 valeurs de seuil de similarité, 4 valeurs de facteur d'inflation et 6 largeurs de fenêtre temporelle, produisant un total de 192 configurations.

	2013-2014-2016	2015
Macro-pureté	0,98	0,98
Macro-pureté inverse	0,83	0,98

Tab. 3.2: Macro- et micro-pureté et pureté inverse pour la configuration similarité : 0,2 ; facteur d'inflation : 1,4 ; décroissance temporelle : 7 jours sur l'ensemble des dépêches AFP des années 2013, 2014 et 2016, comparée à celles obtenues sur 2015.

Nous souhaitons sélectionner la combinaison de paramètres la plus performante en gardant à l'esprit que le recouvrement entre les clusters issus de chaque configuration et ceux de référence sera de taille variable. Pour cette raison, nous faisons primer la macro-pureté sur la micro-pureté et ne cherchons pas nécessairement à minimiser la différence entre pureté et pureté inverse. En privilégiant ainsi les configurations aux puretés les plus élevées en faisant de l'équilibre avec la pureté inverse un critère d'arbitrage secondaire, nous préférons retenir une configuration produisant un clustering de granularité potentiellement plus fine que celui de référence, ce qui nous semble raisonnable compte tenu de nos objectifs pour cette étape.

Le grand nombre de combinaisons rend difficile la visualisation de l'intégralité des performances mesurées. C'est pourquoi la figure 3.8 ne présente que les 15 candidates les plus intéressantes. On peut voir que dans le cas de la macro- comme de la micro-pureté, l'augmentation de la pureté inverse se fait au détriment de la pureté, avec un recul plus net de la pureté comparativement à l'augmentation de la pureté inverse. De plus, on constate que les valeurs de macro-pureté s'échelonnent dans un intervalle de valeurs plus restreint que celles de la micro-pureté.

On observe aussi que l'ensemble des mesures représentées ont un seuil de similarité plutôt permissif, de 0,1 ou 0,2 et un facteur d'inflation faible, prenant les valeurs 1,4 ou 2. On remarque aussi que, toutes choses égales par ailleurs, l'augmentation de la fenêtre de décroissance temporelle semble favoriser la pureté inverse, plus nettement dans le cas de la macro-pureté, les valeurs “micro" étant par ailleurs très élevées et très proches. Cet effet est particulièrement visible quand la fenêtre atteint ou dépasse 7 jours.

En suivant les critères de décision définis ci-dessus, nous retenons la configuration similarité : 0,2 ; facteur d'inflation : 1,4 ; décroissance temporelle : 7 jours comme la meilleure. Les performances de cette configuration pour les deux ensembles d'évaluation sont rapportées par le tableau 3.2.

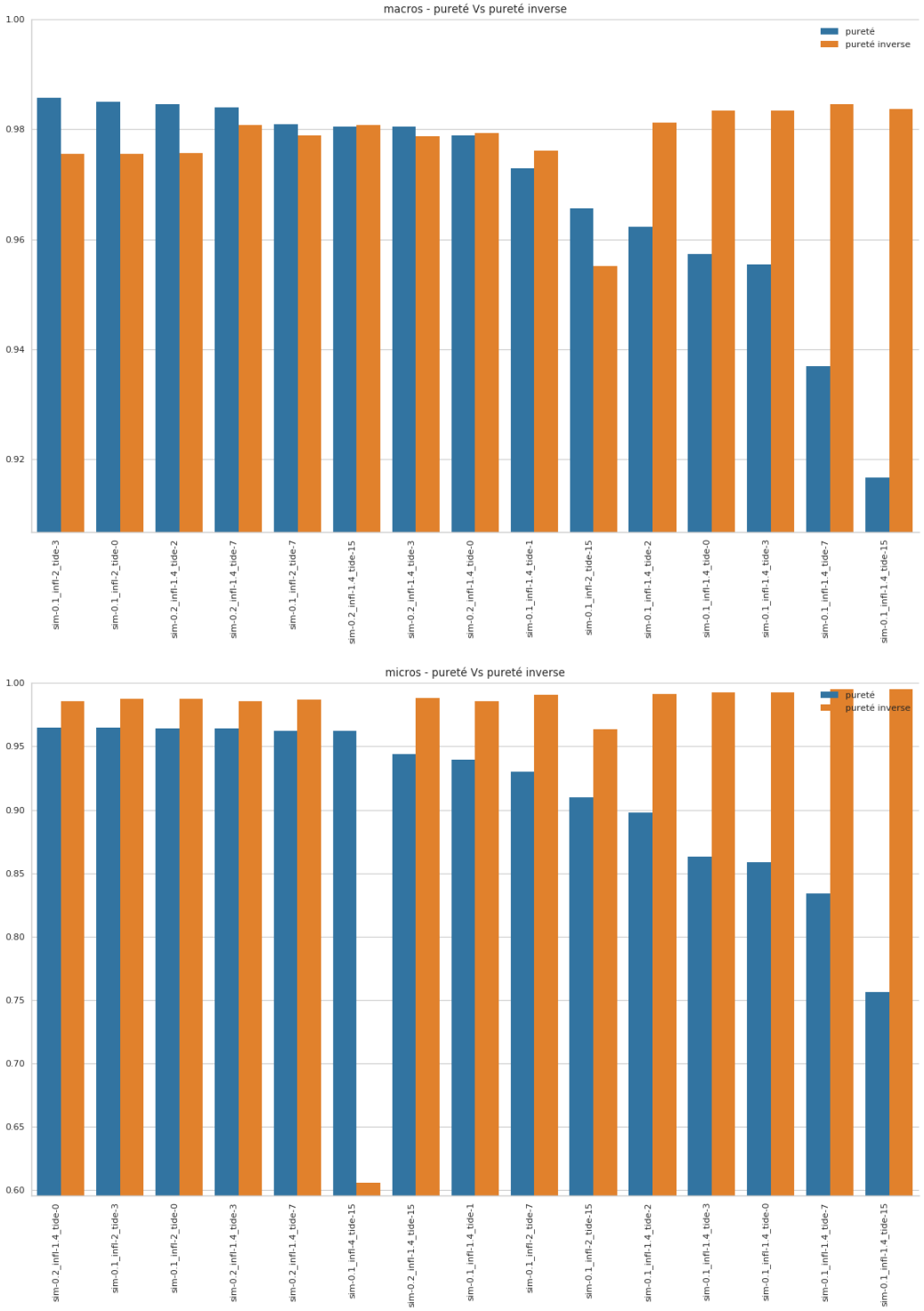


Fig. 3.8: Les 15 configurations présentant la meilleure performance sur les données AFP de l'année 2015. En haut, les valeurs "macros", en bas, les valeurs "micros". Chaque graphe présente la valeur de pureté *vs* celles de pureté inverse. En faisant primer le "macro" sur le "micro" et la pureté sur la pureté inverse, nous retenons la combinaison de paramètres suivante pour la suite de notre travail : similarité : 0,2 ; facteur d'inflation : 1,4 ; décroissance temporelle : 7 jours (sur la figure : `sim-0.2_infl-1.4_tide-7`).

3.5.2 Agrégation d'articles issus du Web autour d'amorces AFP

Méthodologie

L'agrégation d'articles Web ne nécessite l'optimisation que d'un seul paramètre, à savoir le seuil de similarité en-dessous duquel un document Web n'est pas conservé dans le cluster hybride. Nous arrêtons notre choix de manière essentiellement quantitative en comparant les quantités de documents conservées pour chaque valeur candidate.

Pour sélectionner un seuil de similarité adapté, nous avons effectué le filtrage décrit en section 3.3 pour les valeurs de similarité 0,1, 0,2 et 0,3. Nous n'avons pas testé de valeurs supérieures, qui auraient fortement réduit le rappel, sachant que notre objectif est avant tout de réduire le nombre de documents hors sujet dans chaque cluster.

Comme à l'étape précédente, nous effectuons la sélection de la valeur de seuil de similarité optimale sur les clusters hybrides de l'année 2015 et appliquons ce paramètre sur ceux issus du traitement des années 2013-2014-2016.

Résultats et discussion

La figure 3.9 présente le nombre moyen de documents dans chaque cluster résultant de ce filtrage grossier. On constate d'abord une forte dispersion, due au décalage de recouvrement entre les corpus AFP et Web évoqué en section 3.5.

L'importante chute du nombre moyen de documents agrégés entre les valeurs de seuil et en particulier entre les valeurs 0,1 et 0,2 peut s'expliquer quant à elle en partie par la différence stylistique entre les deux types de contenus. Nous avons discuté plus haut des contraintes qui guident la rédaction des dépêches AFP, notamment leurs chapeaux. Bien que l'objectif de cette partie (à savoir informer le lecteur de l'essentiel de la nouvelle développée dans la suite du document) soit le même pour une dépêche d'agence ou un article de presse généraliste, cette dernière dispose de plus de latitude dans sa rédaction (ce qui, rappelons-le, motive notre recours à cette donnée). Néanmoins, ce degré de liberté supplémentaire peut également se traduire par un décalage significatif de traitement de la nouvelle, comme l'illustre la figure 3.10.

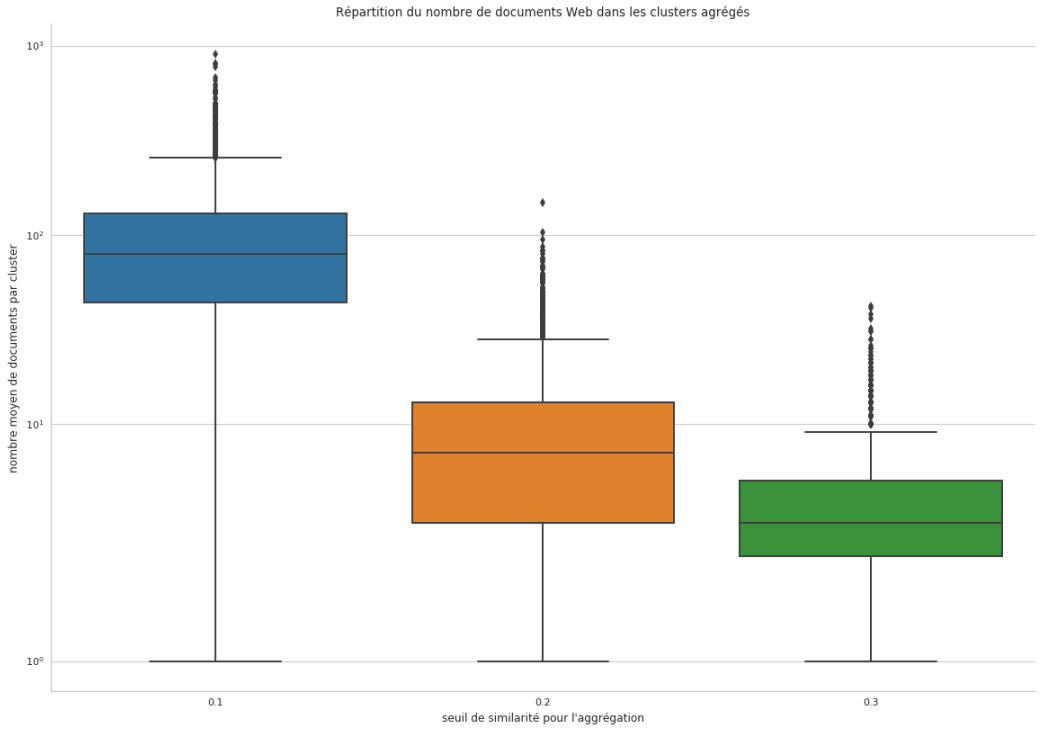


Fig. 3.9: Boîtes à moustaches montrant la répartition du nombre de documents Web dans les clusters agrégés pour différents seuils de similarité, en échelle logarithmique. Compte tenu de la forte dispersion présente pour chaque seuil, de la baisse drastique du nombre de documents agrégés dès la valeur 0,2 et sachant que nous souhaitons privilégier un fort rappel, nous choisissons le seuil 0,1.

AFP

Wildfires wreak havoc across northern California, killing one.

Winds whipped new life Monday into northern California wildfires that have killed a woman, hospitalized firefighters, and reduced hundreds of homes to smoldering ruins.

Web

What It Was Like Inside California's Raging Valley Fire.

Raging fires spread throughout Middletown, California, Sunday, and while most of the residents had fled for their safety, the rescue crews were left looking for unexpected hazards as they fought the flames.

Fig. 3.10: Comparaison des différences pouvant apparaître dans le traitement d'un événement entre l'AFP et une source Web. La dépêche, en haut et en bleu, se focalise sur un traitement factuel tandis que l'article Web, en bas et en vert, adopte un angle plus narratif et centré sur les conséquences pour les victimes.

Six French have passports confiscated for planning Syria trip : security source. *February 23rd*	**Police shooting prompts protests, tear gas in St. Louis.** *August 20th*
Cyclone-hit Vanuatu declares state emergency *March 15th*	**Facebook pledges to combat racism on German platform.** *September 14th*
Football : Three Greek internationals injured in Budapest car crash. *March 30th*	**EU struggles to reach deal on refugee quotas.** *September 14th*
Magnitude 7.5 earthquake hits Nepal : USGS. *April 25th*	**Athletics : Russia provisionally suspended by IAAF.** *November 13th*
Bangladesh upholds opposition politician's death sentence. *July 7th*	**Nepal facing 'medical crisis' as supplies run short.** *November 20th*
Truck bomb kills at least 33 in Baghdad market. *August 13th*	**Sydney lashed by severe storm.** *December 16th*

Tab. 3.3: Titres des références AFP associées aux 12 clusters hybrides annotés pour l'évaluation de l'étape finale de filtrage.

3.5.3 Filtrage du contenu Web

Méthodologie

L'évaluation de cette dernière étape s'est faite sur un jeu d'articles Web annotés par nos soins. Pour ce faire, nous avons sélectionné 12 clusters hybrides associés à 12 dates remarquables dans la liste des événements référencés par Wikipédia pour l'année 2015 [3] et partageant une couverture raisonnable à la fois dans le corpus AFP et dans le corpus Web. Nous avons utilisé les dates comme repères, sans nous contraindre à retenir le cluster hybride correspondant à l'une des instances décrites dans Wikipédia. Le tableau 3.3 montre les titres des références AFP associées à chaque cluster hybride de cette sélection.

La sélection de documents à annoter au niveau des clusters hybrides présente deux avantages. D'abord, elle permet de s'assurer que l'instance est bien représentée à la fois dans le corpus AFP et dans le corpus Web, la couverture de ces deux corpus pouvant fortement différer. En effet, comme on peut le voir dans le tableau en annexe B, nos sources Web sont issues de la presse anglo-saxonne, plus particulièrement des États-Unis, alors que l'AFP est une agence de presse française qui, si elle couvre de larges pans de l'actualité internationale en langue anglaise, traite plus finement les nouvelles européennes et françaises. De ce fait, certains événements peuvent se trouver couverts plus en profondeur dans le corpus Web, conduisant à des ensembles de documents agrégés de tailles inégales, voire en une absence de couverture si l'événement n'a pas été couvert par l'AFP. Réciproquement, un cluster AFP n'ayant pas trouvé de correspondance dans le corpus Web ne sera pas conservé. Les clusters

3. `https://en.wikipedia.org/wiki/2015`.

Magnitude 7.5 earthquake hits Nepal : USGS. A powerful 7.5 magnitude earthquake struck Nepal on Saturday, the United States Geological Survey said, with strong tremors felt across the Himalayan nation and parts of India.
101 dead as 7.8 quake hits Nepal, causing big damage. A powerful earthquake struck Nepal Saturday, killing at least 71 people as the violently shaking earth, collapsed houses, leveled centuries-old temples and triggered avalanches in the Himalayas.
Nepal quake toll reaches 688 : government. KATHMANDU (Reuters) - The death toll from a powerful earthquake that struck Nepal on Saturday has risen to 688, a senior home ministry official told Reuters, with 181 people killed in the capital Kathmandu.

Fig. 3.11: Exemples de relations “forte" et “distante" à une mention de référence. Le texte en haut, en bleu, est la référence ; le deuxième, en vert est considéré comme ayant une relation “forte" à cette dernière tandis que le troisième, en orange représente une relation “distante".

hybrides ne décrivent donc, par construction, que des instances présentes dans les deux corpus.

Le second avantage est de pouvoir utiliser ces clusters hybrides comme baseline à laquelle nous comparerons les résultats après l’opération de filtrage afin de quantifier son efficacité.

Nous avons ensuite sélectionné, pour chaque cluster hybride retenu, les 50 articles Web aux scores d’alignement les plus élevés à l’issue de l’étape de filtrage. Chaque article a ensuite été annoté par trois annotateurs, qui devaient lui attribuer une valeur de pertinence à trois niveaux vis-à-vis de la référence AFP :

— 0 : l’article n’est pas lié à l’instance de référence considérée ;

— 1 : l’article a une relation “distante" à l’instance de référence considérée ;

— 2 : l’article a une relation “forte" à l’instance de référence considérée.

Nous définissons la relation d’un article à la référence AFP comme “forte" si l’article considéré relate strictement les mêmes faits, les variations se situant uniquement au niveau rédactionnel ou dans l’absence ou l’ajout d’informations complémentaires à sa description. La relation “distante" caractérise quant à elle les articles relayant une information périphérique à l’événement de référence mais si spécifique et si intrinsèquement liée à l’événement de référence que la mention de ce dernier est cruciale à la compréhension de l’article. Cette différence est explicitée à la figure 3.11.

Cette distinction a été introduite pour mieux caractériser certains phénomènes liés au développement d’un événement à mesure qu’il se diffuse dans les canaux journalistiques. Deux exemples typiques sont les mises à jour de bilans (essentiellement des bilans de victimes mais pas seulement) et les réactions officielles à des événements, par exemple les revendications d’attentats. Dans ces deux cas, l’événement causal (catastrophe naturelle ou attaque terroriste) est le centre des premiers articles (ou

Paires d'annotateurs	κ de Cohen
(1, 2)	0,67
(1, 3)	0,60
(2, 3)	0,63

κ de Fleiss (tous les annotateurs)
0,61

Fig. 3.12: Accords inter-annotateurs mesurés par le kappa de Cohen et le kappa de Fleiss. Ils convergent autour de 0,61, une valeur modérée mais satisfaisante compte tenu du caractère parfois épineux de la tâche d'annotation.

dépêches) publiés, chronologiquement. Par la suite, l'évolution du bilan comme la revendication de l'attaque déclenchent la production de flux d'informations indépendant des événements causaux. Pourtant dans un cas comme dans l'autre, la mention de l'événement causal est une condition nécessaire à la compréhension, et même à la pertinence de l'information relayée.

Ce processus d'annotation s'est avéré parfois ardu, révélant toute la complexité de la définition d'un événement et de ses frontières. Nous avons établi et fixé certaines conventions dans un guide d'annotation, accessible au lecteur en annexe A, afin de produire une annotation à la fois rapide et cohérente entre les annotateurs et reproductible dans la perspective éventuelle d'une phase d'annotation ultérieure. Le tableau 3.12 rapporte la moyenne de l'accord inter-annotateur pour chaque paire d'annotateurs (moyenne des kappas de Cohen sur l'ensemble des instances pour chaque paire d'annotateurs) ainsi que pour l'ensemble des annotateurs (moyenne des kappas de Fleiss sur les instances). Deux tableaux détaillant l'intégralité des kappas de Cohen et de Fleiss sont donnés en annexe C. On constate que les deux mesures convergent autour d'une valeur de 0,61, ce qui est généralement considéré comme un accord modéré. Ce résultat est selon nous un reflet des difficultés évoquées plus haut. Bien que ces valeurs de kappas conduisent à nuancer les résultats de l'évaluation, ces derniers restent de notre point de vue satisfaisants et exploitables.

Nous mettons en œuvre ce schéma d'annotation au travers de deux modalités d'évaluation. Dans la première modalité, dite "stricte", seuls les documents étiquetés comme ayant une relation "forte" sont pris en compte dans les exemples positifs. La seconde modalité, dite "large", y inclut également les documents étiquetés "distants".

Nous souhaitons évaluer la qualité du filtrage produit à l'issue de cette étape selon deux points de vue : d'une part celui des articles Web qui composent le cluster avant et après ce filtrage ; d'autre part celui de la sélection d'articles résultant du calcul du coude de la distribution des scores. Nous mesurerons les performances de chacune de ces modalités à l'aide d'un jeu de métriques différentes, reprenant celles utilisées par GLAVAŠ et ŠNAJDER (2013), que nous appliquons sur les ensembles

d'articles (complets ou sélection par le coude) ordonnés par ordre décroissant de leurs similarités respectives (cosinus ou par alignement).

Sur l'ensemble complet des articles avant et après filtrage, nous calculons la *Mean Average Precision* (MAP) et la R-précision, définies respectivement comme suit :

$$AP = \frac{\sum_{k=1}^{N} (P(k) * rel(k))}{R}$$

où $k =$ rang de l'article, $N =$ nombre total d'articles retournés, $P(k) =$ précision du sous-ensemble d'articles au rang k, $rel(k) = 1$ si l'article k est pertinent, 0 sinon et $R =$ nombre total d'articles pertinents à retrouver.

$$R\text{-}préc. = \frac{r}{R}$$

où $r =$ nombre d'articles pertinents récupérés parmi les R premiers résultats.

Le sous-ensemble formé par les articles filtrés par le coude de la distribution des scores d'alignement est évalué par le calcul de la précision, du rappel et de la F-mesure.

Résultats et discussion

L'ensemble des performances est résumé dans le tableau 3.13. Nous indiquons les résultats globaux (colonnes "total") ainsi que le détail des performances en fonction de l'accord inter-annotateur pour les instances (colonnes "κ fort" pour les instances ayant un accord inter-annotateur supérieur à 0,61, "κ faible" pour les instances ayant un accord inter-annotateur inférieur à ce seuil), ce qui permet dans un premier temps de constater que les instances ayant un κ faible dégradent fortement l'ensemble des performances en valeur absolue, ce qui n'est pas surprenant mais apporte un degré de nuance qu'il nous a semblé important de mentionner.

Si l'on compare le passage de la modalité "stricte" à la modalité "large", on peut noter une nette augmentation des performances sur les mesures avant application du critère du coude (colonnes "total", lignes "MAP" et "R-préc."), ce qui témoigne de la difficulté de la tâche : la frontière d'un événement étant difficile à définir clairement, notamment en raison de la proximité du contenu des articles, le système tend à produire plus de faux positifs dans la configuration la plus stricte. Cette augmentation se retrouve également dans les résultats des mesures après application du coude (colonnes "total", lignes "F-mesure") mais uniquement après le filtrage des clusters.

Stricte	**Avant filtrage**			**Après filtrage**		
	κ fort	κ faible	total	κ fort	κ faible	total
MAP	91	49,7	73,8	97,7	35,8	71,9
R-préc.	87,3	44,5	69,5	95,3	24,5	65,8
Précision	90,4	60	77,7	70,5	16,8	48,1
Rappel	60,3	27,9	46,8	95,8	77,7	88,3
F-mesure	67,2	25,8	50	76,9	25,9	55,7
Large	**Avant filtrage**			**Après filtrage**		
	κ fort	κ faible	total	κ fort	κ faible	total
MAP	90,7	67,1	80,9	97	65	83,7
R-préc.	84,9	67,2	77,5	93	50.3	75,2
Précision	90,4	70,7	82,2	73,4	52,2	64,5
Rappel	57,7	12,6	38,9	95,5	94.5	95,1
F-mesure	65,4	15,2	44,5	79	62,3	72

Fig. 3.13: Les résultats des différentes modalités d'évaluation du clustering en instances sur les 12 instances annotées. Le tableau supérieur synthétise la modalité "stricte" et en dessous la modalité "large". Pour chaque modalité, "κ fort" indique un accord inter-annotateur supérieur ou égal à 0,61 (7 instances), "κ faible" un accord inter-annotateur inférieur à 0,61 (5 instances) et "total" indique les résultats pour l'ensemble des instances réunies.

La dégradation des performances mesurées pour les clusters non-filtrés laisse à penser que le critère du coude tendrait à y être très sélectif, ce qui le défavorise dans le cas d'une évaluation plus souple. Cette interprétation est renforcée par le fait que cette baisse de la F-mesure est causée par celle du rappel, alors que la précision progresse.

Si l'on s'intéresse maintenant à l'apport du filtrage des clusters (comparaison des colonnes "total" pour chaque modalité), on peut observer qu'il dégrade légèrement les performances avant le critère du coude (MAP et R-précision), mais améliore fortement les résultats après application de ce dernier (Précision, Rappel, F-mesure), en diminuant la précision mais en améliorant nettement le rappel, résultant en des performances supérieures en termes de F-mesure. Nous concluons donc que cette étape possède une réelle plus-value en remplissant son objectif de consolidation des descriptions d'instances constituées par l'agrégation d'articles Web autour de dépêches AFP.

3.6 Conclusion

À l'issue de cette première phase, nous obtenons des instances d'événements sous la forme de clusters de dépêches AFP et d'articles Web (que nous nommons descriptions

d'instances), formées en prenant en compte leurs caractéristiques temporelles et lexicales. Chaque description d'instance se compose d'une dépêche AFP dite référence (pensée comme un ancrage chronologique), d'un groupe (cluster) de mentions textuelles décrivant l'instance et d'une matrice d'alignement de termes. Dans la phase suivante de nos travaux, nous exploiterons ces informations pour rapprocher différentes descriptions d'instances les unes des autres dans le but de constituer des groupes s'apparentant à des types d'événements.

4 De l'instance au type

> (Bohort) - *[. . .] Attila, vous le comptez comme un Hun ?*
> (Arthur) - *Ben évidemment j'le compte pas comme un Ibère !*
>
> — **Nicolas Gabion et Alexandre Astier**
> Kaamelott, Livre III, Le fléau de Dieu II

Ce chapitre détaille la deuxième phase de notre travail, qui vise à exploiter les descriptions d'instances d'événements obtenues lors de la première phase, c'est-à-dire des groupes de documents composés d'une dépêches AFP servant de référence chronologique, d'un cluster de mentions textuelles combinant des dépêches AFP à des articles Web et d'une matrice d'alignement de mots. Notre objectif est de regrouper les instances ainsi décrites en types d'événements. Cette phase d'identification de types d'événements se rapproche des problématiques de la tâche *topic detection* de la campagne TDT et a également été explorée par des travaux connexes à cette campagne tels que (Ferret, 1998) ou (Filatova et al., 2006). Elle peut aussi être vue comme une tâche effectuée de façon implicite par la plupart des systèmes induisant des représentations génériques d'événements présentés en section 2.2.3, que nous préférons traiter de manière isolée et explicite car elle présente à notre avis une complexité et des enjeux qui le justifie. De plus, nous pensons qu'une structuration en types d'événements constitue un meilleur point de départ pour l'induction de schémas que les instances, bien que certaines approches aient abordé le problème sous d'autres angles (par exemple, Ahn (2017) induit des types d'événements à partir de rôles identifiés dans des mentions).

Cette phase comporte deux étapes :

1. la mise en place d'une mesure de similarité adaptée à la comparaison des descriptions d'instances sur un plan thématique ;
2. l'identification de types d'événements par un clustering des descriptions d'instances similaires.

Elle est synthétisée graphiquement en figure 4.1. Nous reproduisons ici, dans une certaine mesure, notre approche d'identification d'instances au sein du corpus AFP, en faisant reposer le processus d'induction sur une étape de clustering, elle-même

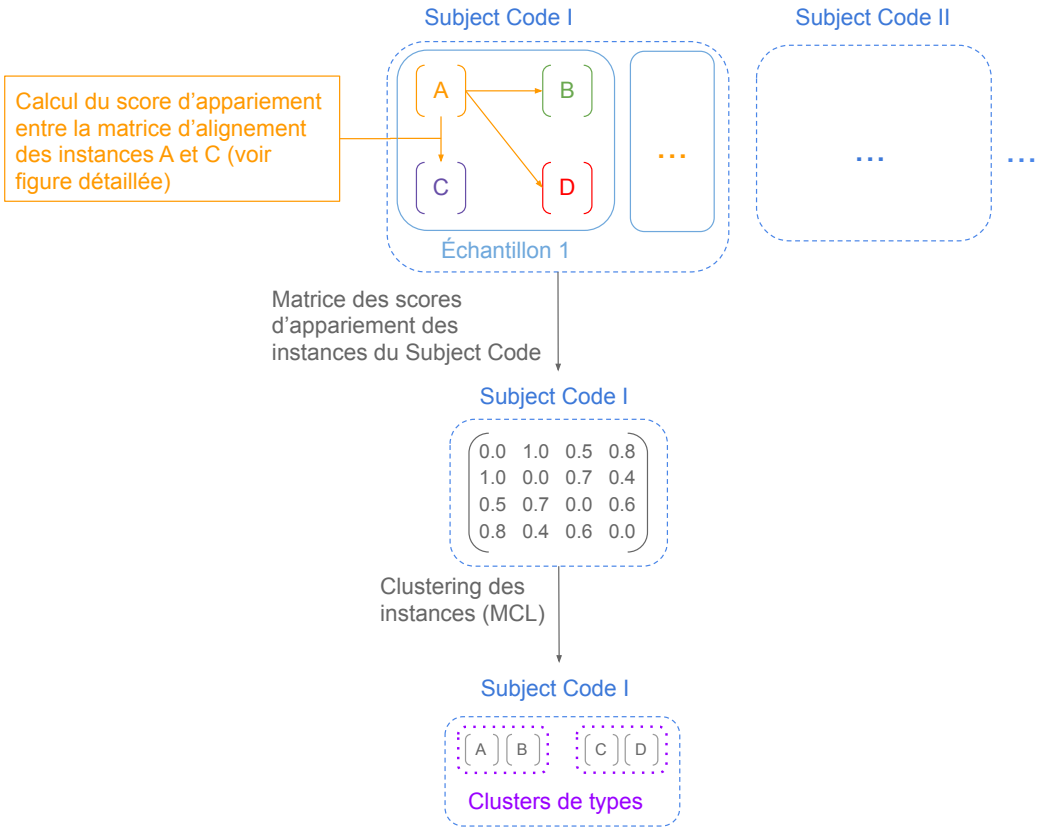


Fig. 4.1: Synthèse des opérations aboutissant à la création de clusters d'instances groupés par types d'événements. Les opérations de découpage par Subject Code ainsi que d'échantillonnage des instances sont détaillées en section 4.1. Le détail du calcul du score d'appariement entre deux instances est donné en figure 4.2.

fondée sur une mesure de similarité dont le calcul tient compte des spécificités propres à la granularité des événements considérés. Nous faisons ce choix car nous pensons que la tâche d'identification de types d'événements présente des caractéristiques similaires, à savoir de nécessiter une certaine flexibilité à la fois sur le nombre de clusters à produire et sur le critère de constitution de ces clusters, deux points que nous avons pu exprimer de manière satisfaisante, selon nous, dans la phase précédente.

4.1 Découpage thématique et échantillonnage des instances

Comme nous le verrons dans la section suivante, la mesure de similarité entre instances d'événements que nous mettons en place a pour inconvénient une grande complexité combinatoire. Afin de maîtriser cette complexité, nous avons dans un

premier temps découpé le corpus de descriptions d'instances de façon à créer des sous-ensembles thématisés.

Pour ce faire, nous exploitons le fait que chaque dépêche AFP est décrite par au moins un, et le plus souvent plusieurs, Subject Codes IPTC. La taxonomie des Subject Codes comprenant plusieurs niveaux hiérarchiques, si une dépêche est associée à un Subject Code de bas niveau, elle portera également les Subject Codes des niveaux supérieurs. Une dépêche peut par ailleurs être associée à plusieurs Subject Codes de bas niveaux, ce qui en fait une forme de *soft clustering*. Nous répartissons chaque description d'instance en fonction des Subject Codes associés à sa référence AFP et appartenant à la liste de 72 Subject Codes thématiques que nous avons définis. Nous obtenons ainsi des sous-ensembles thématiques regroupant des descriptions d'instances, chacune pouvant se trouver dans différents sous-ensembles. Ce découpage n'a pas été appliqué lors de la phase précédente car la multiplicité des Subject Codes disponibles pour chaque dépêche en fait une donnée délicate à exploiter à ce niveau. De plus, nous pensons que la formation des instances peut s'effectuer principalement sur un critère sémantique. À l'inverse, introduire une dimension thématique prend du sens pour l'identification de types car notre critère de regroupement doit dépasser la notion de paraphrase.

À l'issue d'une première expérimentation, il s'est avéré que ce découpage n'était pas suffisant. En effet, comme on peut le voir dans le tableau 4.1, la distribution du nombre d'instances pour chaque Subject Code peut varier en atteignant plusieurs centaines pour certains d'entre eux. Chacun de ces cas extrêmes, bien que rare, constitue à lui seul un cas trop complexe du point de vue combinatoire. Nous tenons néanmoins à rester en mesure d'exploiter au mieux les informations à notre disposition et souhaitons donc mettre en place une stratégie visant à maximiser le nombre d'instances traitées. Pour ce faire, nous procédons par sous-échantillonnage, c'est-à-dire que nous subdivisons le contenu des Subject Codes comportant plus de 100 descriptions d'instances en plusieurs échantillons de 100 instances sélectionnées au hasard. Cette valeur de 100 instances a été choisie en examinant la distribution du nombre d'instances parmi les Subject Codes : une grande majorité d'entre eux n'en contiennent pas plus de 100. Cet échantillonnage restera donc cantonné aux cas les plus problématiques. De plus, la complexité combinatoire dans le cas d'un ensemble de 100 instances est raisonnable et permet de travailler et d'expérimenter à grande échelle sans temps de traitement prohibitifs. Cet échantillonnage suppose néanmoins que la distribution des phénomènes d'intérêt pour la caractérisation de types d'événement à partir de nos descriptions d'instances soit uniforme entre les descriptions. Il implique par ailleurs la mise en place d'un traitement intermédiaire lors de l'induction des types d'événements à partir des descriptions d'instances, que nous détaillerons dans le chapitre suivant.

Nombre de descriptions d'instances par Subject Code	Nombre de Subject Codes 2015	2013-2014-2016
100 ou moins	53	45
101 à 200	8	12
201 à 300	3	4
301 à 400	3	4
401 à 500	0	0
501 à 600	1	2
601 à 700	2	0
701 à 800	0	1
plus de 800	2	3
Total	**72**	**71**

Tab. 4.1: Distribution du nombre d'instances identifiées pour chaque Subject Code IPTC à partir du Subject Code associé à la référence AFP de chaque instance, pour les années 2013 à 2016. Les Subject Codes sont classés par ordre croissant du nombre d'instances pour l'année 2015, utilisée pour la mise au point du système.

À l'issue de ces deux étapes, nous disposons donc d'un ensemble de matrices de scores synthétisant chacune la similarité entre les instances regroupées sous un même Subject Code. Cette similarité est construite par le calcul d'un score d'appariement des matrices d'alignement associées à chaque description d'instance, lui-même le résultat de la résolution du problème d'affectation appliqué au graphe biparti formé dont les sommets sont formés par les lignes des matrices considérées et les arêtes sont pondérées par la similarité sémantique et syntaxique entre ces ensembles de mots. Une matrice de score est calculée pour chaque échantillon de 100 descriptions d'instances à l'intérieur de chaque Subject Code IPTC. Ce double découpage du corpus d'instances (Subject Codes et échantillons d'instances) permet de maîtriser la grande complexité combinatoire de l'approche présentée tout en structurant le corpus de façon pertinente sans sacrifier au volume de données exploité. En effet, le découpage par Subject Codes respecte des contraintes thématiques et le choix de 100 instances comme valeur d'échantillonnage permet de restreindre l'impact de ce traitement à la minorité de Subject Codes posant ces problèmes calculatoires. Nous conservons également le détail des appariements calculés.

4.2 Appariement de descriptions d'instances

Nous souhaitons définir un score assimilable à une mesure de similarité qui nous permette de rapprocher des descriptions d'instances selon le type d'événement qu'elles décrivent. Pour ce faire, nous exploitons les matrices d'alignement associées à chaque description d'instance issues du filtrage par alignement (cf. 3.4), en faisant l'hypothèse que cette étape les a dotées de certaines propriétés :

— le processus d'alignement sélectionne uniquement les mots les plus caractéristiques de l'instance considérée ;

— il permet également de capturer des formes lexicales différentes mais sémantiquement proches, avec pour résultat que chaque instance est décrite par des ensembles de mots variés mais associés à des rôles sémantiques similaires.

Nous considérons par ailleurs que le score d'alignement et le critère du coude ont permis de définir un sous-ensemble de documents pour lesquels ces propriétés se vérifient avec une bonne fiabilité. Rappelons qu'une matrice d'alignement se compose d'une colonne par document aligné, la première colonne étant associée à la référence AFP du cluster hybride représenté et que chaque ligne correspond à un mot de la référence ayant été aligné dans au moins un document. De plus, une matrice contient les alignements trouvés pour le titre de la référence (avec le titre de chaque document) et pour son chapeau (avec le chapeau de chaque document).

Nous définissons notre score de similarité entre les instances en adaptant celui sous-tendant l'algorithme de SULTAN et al. (2014) pour mesurer la compatibilité de deux mots candidats lors de l'alignement de deux phrases. Ce score est scindé en deux composantes, l'une sémantique, l'autre syntaxique. La composante sémantique vaut 1 si les deux mots comparés sont identiques ou partagent la même étiquette d'entité nommée, 0,9 s'ils sont appairés dans la base PPDB, 0 sinon. La composante syntaxique se focalise sur les mots du contexte phrastique liés aux candidats par des dépendances syntaxiques et recherche celles appartenant à un ensemble prédéfini de relations équivalentes en tenant également compte de leur orientation respective et de l'étiquette en partie du discours associée au mot de contexte considéré. Le score de similarité entre les deux mots candidats t_1 et t_2 est alors donné par :

$$sim(t_1, t_2) = \theta * sem\_score + (1 - \theta) * syn\_score$$

où $\theta = 0,9$ (valeur retenue par les auteurs). Précisons que, comme lors de la phase précédente, les mots que nous manipulons associent chacun le lemme d'une forme de surface et son étiquette en partie du discours. Considérant que chaque instance d'événement est décrite par une matrice dont chaque ligne représente des mots ayant fait l'objet d'un alignement, nous appliquons ce calcul à toutes les paires de mots formées dans toutes les paires de lignes des matrices d'alignement associées aux instances que l'on souhaite apparier, représentées par les flèches noires en partie supérieure de la figure 4.2. La somme de ces scores individuels permet d'obtenir un score d'appariement pour la paire de ligne. Ce score d'appariement est ensuite calculé pour toutes les paires de lignes. Ainsi, dans la figure 4.2, le score de similarité décrit sera calculé entre le mot *Several* et le mot *Paris*, trois fois, puis entre *multiple* et *Paris*, trois fois encore, avant que l'ensemble de ces scores soit sommé pour donner le score entre les deux lignes. Puis la même procédure sera répétée jusqu'à atteindre les paires *Several/casualties*, *Several/fatalities*, *Several/dead*,

multiple/casualties, *multiple/fatalities* et *multiple/dead* dont la somme des scores formera le score de la paire de lignes considérées. On appliquera ensuite la même opération avec la ligne suivante, composée des mots *people* et *civilian*.

Intuitivement, on souhaite qu'une ligne de la matrice d'une instance donnée ne puisse s'appairer de façon pertinente qu'avec au plus une ligne de la matrice de l'autre instance considérée. On devine dès lors qu'appliquée dans sa forme la plus directe, l'approche décrite ci-dessus consacrera beaucoup de ressources à des appariements non pertinents. Afin de réduire ce problème, nous restreignons le calcul complet du score d'appariement de deux lignes à celles partageant au moins un mot commun. Si l'on se réfère de nouveau à la figure 4.2, cela signifie qu'aucun calcul ne sera effectué pour les lignes contenant les mots *Several* et *multiple* d'une part et *American*, *US* et *American* d'autre part, d'où leur représentation grisée dans la partie inférieure de la figure. Cette condition semble raisonnable dans la mesure où les lignes d'une matrice peuvent être considérées comme des vocabulaires formés par des sources en partie redondantes (l'AFP). La structure des matrices d'alignement, qui agrègent à la fois les alignements faits entre les titres des documents et ceux faits entre les chapeaux (corps du texte), soulève une difficulté plus subtile. Elle conduit en effet à ce que deux lignes d'une même matrice peuvent comporter le même mot. Ce cas présente l'inconvénient d'aller contre notre contrainte énonçant qu'une ligne d'une matrice donnée ne peut trouver au plus qu'un correspondant pertinent dans une autre matrice. Nous réduisons ce problème en fusionnant au préalable, dans chaque matrice, les lignes dont le mot associé à la dépêche de référence est dupliqué.

Enfin, nous souhaitons que le score d'appariement attribué à une paire de matrices d'alignement donnée maximise la somme des scores d'appariement des lignes qui les composent, sous la contrainte que chaque ligne d'une matrice donnée se voit associée à au plus une ligne de l'autre. Ce problème s'assimile à la recherche d'un couplage parfait de poids maximal dans un graphe biparti, aussi appelé problème d'affectation, dans lequel :

— les sommets sont les lignes des matrices d'alignement (après fusion des doublons) ;

— les arêtes sont les appariements qui ont pu être formés entre les lignes ;

— le poids d'une arête correspond au score d'appariement calculé entre les deux lignes (c'est-à-dire les sommets du graphe) associées.

Dès lors, nous pouvons calculer le score d'appariement des matrices en résolvant le problème d'affectation dans le graphe correspondant. Il s'agit d'un problème classique de recherche opérationnelle qui peut être résolu en temps polynomial grâce à l'algorithme hongrois (KUHN, 1955). Bien que ce problème dans sa forme classique consiste à rechercher un couplage parfait de poids minimum dans un graphe biparti, il suffit d'appliquer l'algorithme sur un graphe dont on a changé au

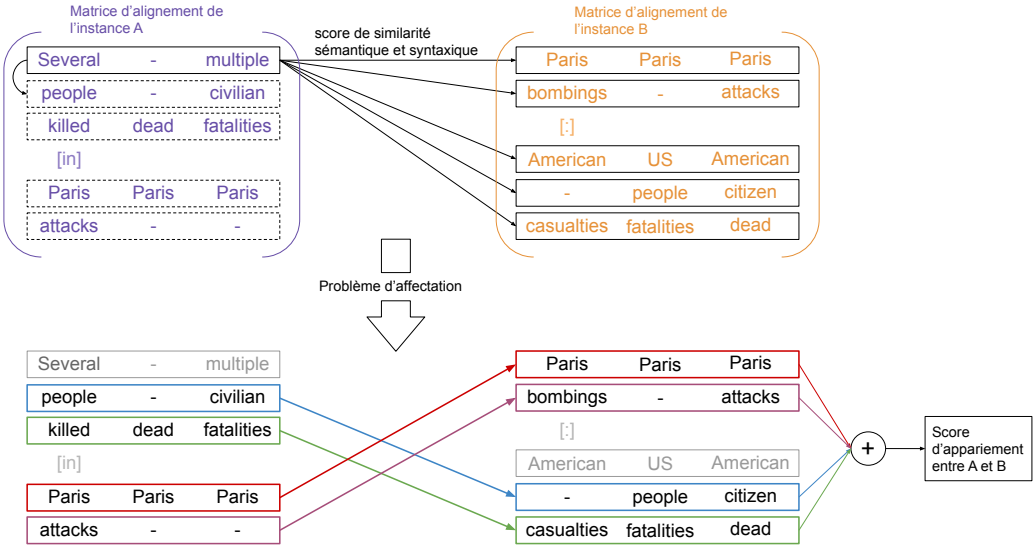


Fig. 4.2: Détail du calcul du score d'appariement entre deux instances d'événements.

préalable le signe des poids pour trouver le couplage parfait de poids maximum. La figure 4.2 illustre en détail l'ensemble du processus, depuis le calcul de similarité entre deux matrices jusqu'au résultat de la résolution du problème d'affectation.

C'est cette méthode qui, en dépit des optimisations réalisées, souffre d'une grande complexité combinatoire, laquelle limite sa mise en œuvre à grande échelle sur l'ensemble des descriptions d'instances disponibles et a motivé la mise en place des stratégies de découpage et d'échantillonnage décrites en section 4.1. Nous pouvons alors appliquer l'approche détaillée ici à chaque échantillon de chaque Subject Code, afin d'obtenir une matrice de similarité des instances contenues dans chacun de ces sous-ensembles.

4.3 Mise en évidence de types par clustering des paires d'instances

Les matrices de similarité entre instances construites à l'étape précédente constituent la base pour construire des regroupements d'instances d'événements assimilables à des types d'événements. Pour effectuer cette construction, nous réutilisons l'algorithme MCL, déjà présenté en section 3.2. Ce choix est en partie motivé par les mêmes raisons exposées alors, à savoir qu'il s'agit d'un algorithme pertinent pour notre objet de travail (c'est-à-dire une matrice de similarité) et qu'il ne nécessite pas qu'un nombre de clusters cible lui soit indiqué *a priori*, ce qui est intéressant dans cette configuration car certains Subject Codes ont une granularité assimilable à des types d'événements (par exemple le Subject Code "tremblement de terre") alors

que d'autres ont une dimension plus thématique (notamment "application de la loi") qui peut recouvrir plusieurs types d'événements différents. Bien que nous ne nous attendions pas à rencontrer un grand nombre de types au sein d'un Subject Code, laisser une certaine flexibilité au système semble souhaitable.

Une fois ces clusters obtenus, nous les enrichissons des résultats liés aux graphes bipartis optimisés. La description des instances constituant un cluster est détaillée par l'ensemble des mots appairés et des scores associés et ce, pour toutes les paires d'instances du cluster, afin de disposer de ces informations pour la suite de notre travail.

Enfin, nous voulons exploiter au mieux les informations à notre disposition dans les Subject Codes ayant nécessité un échantillonnage. Pour cela, nous fusionnons les clusters (c'est-à-dire les types d'événements induits) issus de différents échantillons du même Subject Code (cf. 4.1) en exploitant leurs mots communs selon une approche gloutonne. Un cluster d'un échantillon donné est fusionné avec le cluster de chaque autre échantillon avec lequel il a le plus de recouvrement. La figure 4.3 illustre ce processus. Nous proposons trois variations de cette notion de recouvrement :

— fusion par recouvrement lexical : la similarité entre deux clusters-types est ici mesurée par la proportion de mots uniques en commun ;
— fusion par recouvrement fréquentiel : la proportion de mots uniques en commun entre deux clusters-types est pondérée par la fréquence d'apparition des mots considérés dans leurs clusters respectifs ;
— fusion par similarité : la proportion de mots uniques en commun entre deux clusters-types est cette fois pondérée par le score d'appariement calculé au préalable.

Pour résumer cette étape, nous identifions des types d'événements à partir de la mesure de similarité entre instances établies à l'étape précédente (décrite en section 4.2). Le calcul de cette mesure induit une forte complexité combinatoire impliquant l'échantillonnage de notre corpus dont nous devons tenir compte à cette étape. Pour ce faire, Nous appliquons dans un premier temps l'algorithme de clustering déjà utilisé pour l'induction d'instances à partir de mentions sur chacun de ces échantillons, avant de mettre en place une stratégie de fusion des résultats visant à exploiter le plus d'informations disponibles.

4.4 Évaluation

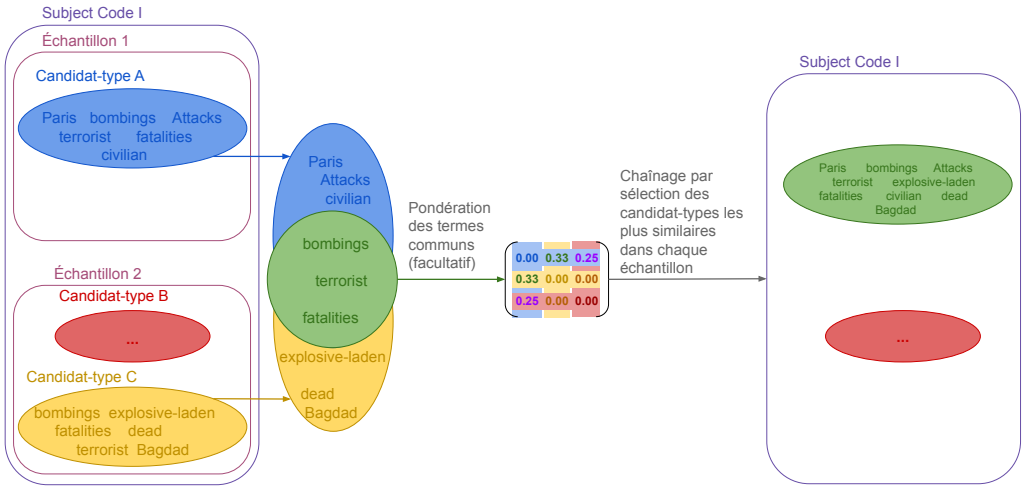


Fig. 4.3: Résumé de la stratégie de fusion des clusters candidats-types issus de différents échantillons d'un même Subject Code. On remarquera que la similarité entre des candidats-types du même échantillon n'est pas calculée.

Fig. 4.4: Détail du dendrogramme résultant du clustering hiérarchique des documents AFP de la correspondance d'instance. On peut voir que les WETs "LODLAM summit", "summit", "BRICS summit", "G7 summit", "G8 summit", "G20 summit", "NATO summit", à la base du clustering, sont regroupés dans un cluster de plus haut niveau avec une homogénéité qui nous permet de l'envisager comme un type "sommet politique".

4.4.1 Présentation des ressources d'évaluation

À partir de la correspondance entre dépêches AFP et instances Wikidata présentée en section 3.5.1, RUDNIK et al. (2019) dérivent une correspondance supplémentaire liée à la notion de type d'événement et associant une dépêche à un Type d'Événement Wikidata (abrégé WET pour *Wikidata Event Type*). On rappelle que le WET désigne, pour chaque instance Wikidata, la valeur de son champ `instance_of`. Cette approche reste néanmoins d'une granularité très fine, ce qui a poussé les auteurs à mettre en œuvre une approche de clustering visant à agréger ces WET et leurs descriptions en types plus généraux. C'est le résultat de ce travail que nous utilisons pour notre évaluation. Un extrait du résultat de ce clustering est donné en figure 4.4.

4.4.2 Présentation de la baseline d'induction de types

Nous souhaitons mettre en perspective les résultats bruts de l'évaluation de notre méthode d'induction de types d'événements avec la ressource décrite en section précédente, notamment afin de caractériser la contribution de notre approche d'hybridation avec des articles Web (agrégation + filtrage). Pour ce faire, nous appliquons également notre protocole d'évaluation (détaillé dans la section suivante) à une approche baseline consistant à :

— segmenter les clusters AFP (issus de la première étape de la première phase, décrite en section 3.2) par les Subject Codes de leurs références chronologiques respectives. Chaque cluster AFP représente ainsi une description d'instance, à l'image des résultats de notre méthode après l'étape de filtrage par alignement ;

— construire une représentation de chaque cluster sous la forme d'un vecteur sac de mots binaire ;

— calculer la matrice de similarité cosinus de ces instances AFP à l'aide de l'algorithme APSS ;

— appliquer le MCL sur cette matrice. Chaque cluster résultant sera considéré comme un type d'événement et évalué individuellement.

Cette approche simple repose sur l'utilisation directe des clusters AFP, sans les faire passer par les étapes d'agrégation et de filtrage. L'induction des types d'événements reproduit la même méthode que celle du clustering AFP, à l'exception de la composante temporelle (c'est-à-dire la décroissance exponentielle du score de similarité en fonction du temps), qui n'a pas lieu d'être dans la perspective de faire émerger des types d'événements. Le processus est par conséquent régi par deux paramètres : le seuil de similarité de l'APSS et le facteur d'inflation du MCL. Nous reprenons les ensembles de valeurs explorés pour ces paramètres lors du clustering AFP et considérons toutes leurs combinaisons (recherche exhaustive). Nous n'optimisons pas conjointement le clustering en instance et en type car cela n'a pas été fait pour notre méthode.

4.4.3 Méthodologie

Nous utilisons les mesures de pureté et de pureté inverse pour étudier la qualité des clusters d'instances obtenus (qui nous serviront par la suite de descriptions de types d'événements), de la même façon que nous l'avons fait pour l'évaluation du clustering AFP (au niveau instance). Comme dans les évaluations précédentes, les données de l'année 2015 servent à l'exploration et la sélection d'une configuration que nous appliquons ensuite sur l'ensemble des années 2013, 2014 et 2016.

De plus, nous n'avons pas mené d'expérimentations visant à optimiser le facteur d'inflation du MCL car les essais préliminaires réalisés lors du développement de la méthode nous ont montré qu'aucune des valeurs recommandées par l'auteur ne produisait plus d'un cluster par Subject Code, à l'exception de la valeur maximale, c'est-à-dire 6 (rappelons que plus le facteur d'inflation est élevé, plus l'algorithme a tendance à produire des clusters de petite taille). Ne souhaitant par ailleurs pas explorer de valeurs en dehors de celles recommandées, nous avons décidé de nous cantonner *de facto* à celle-ci.

Par ailleurs, nous évaluons séparément les Subject Codes contenant moins et plus de 100 instances, afin de pouvoir analyser plus finement l'influence de nos différentes stratégies d'échantillonnage. Nous comparons également l'impact des trois stratégies de fusion d'échantillons présentées en section 4.3 en y ajoutant une quatrième, dite "sans fusion", pour laquelle nous n'effectuons l'induction de types qu'à partir d'un seul d'entre eux. Cette stratégie a pour but de mieux quantifier la contribution (positive ou négative) des autres stratégies.

Enfin, rappelons que le score de similarité entre les mots des matrices, qui forme la base de toute notre stratégie, est scindé en deux composantes, sémantique et syntaxique. Dans la forme originelle utilisée par SULTAN et al. (2014), la composante syntaxique (liée au contexte des candidats) n'est calculée que si un score de similarité sémantique non nul a été préalablement trouvé entre les candidats. Comme nous l'avons vu, cette similarité sémantique entre deux candidats est établie si les deux formes sont identiques ou appairées dans PPDB [1]. Or à l'échelle de différentes instances, les mots qu'il peut être intéressant de rapprocher peuvent être très éloignés sémantiquement (donc non identiques et non appairés dans PPDB). En revanche leur fonction et leur contexte dans leurs phrases respectives peuvent être similaires, ce qui peut être détecté par les équivalences établies entre dépendances syntaxiques, utilisées pour le calcul de la composante syntaxique.

Imaginons par exemple différentes instances d'événement de type "vol" (au sens du délit) : on souhaiterait être capable d'apparier les objets volés, qu'il s'agisse d'un tableau ou d'une voiture. Or, il sera impossible de le faire sur la base de l'identité des formes ou d'une relation de synonymie. De plus, il n'existe à notre connaissance aucune ressource permettant de lier des mots décrivant des choses qui peuvent être volées. En revanche il est raisonnable de penser qu'en tant que noms communs objets de verbes synonymes de "voler", ils occupent la même fonction dans leurs instances respectives. Notons que cette hypothèse repose en partie sur le fait que nos instances sont groupées de manière thématique (*via* les Subject Codes), ce qui contribue à désambiguïser le sens des verbes et par extension, celui de leurs arguments.

1. Rappelons que PPDB est une ressource qui, entre autres, appaire de nombreux mots synonymes.

En conséquence, nous testerons deux modes de calcul du score de similarité : la première effectuera le calcul de la façon originelle, en conditionnant celui de la composante syntaxique au résultat de la similarité sémantique des candidats, tandis que la seconde calculera systématiquement les deux composantes, sans conditionner l'une à l'autre.

4.4.4 Résultats et discussion

Nous discutons d'abord les résultats de notre baseline, donnés en figure 4.5. En raison de la grande quantité d'informations (nous construisons cette paire de graphiques pour les modalités macro et micro de la pureté), nous ne présentons ici que les résultats associés à la mesure de macro-pureté. Les graphiques correspondant à la micro-pureté sont visibles en annexe D.

Nous nous intéresserons dans un premier temps aux valeurs représentées dans le cartouche blanc de chaque configuration (au début de chaque barre d'histogramme). Elles indiquent le nombre de documents ayant servi à l'évaluation de la configuration (c'est-à-dire le nombre de documents de référence trouvés dans nos résultats) par rapport au nombre maximum de documents disponibles (c'est-à-dire le plus petit des deux ensembles, référence ou clustering de la baseline). On constate que ces deux valeurs diminuent rapidement à mesure que les configurations deviennent plus restrictives, ce qui signifie que les clusterings que nous obtenons pour chaque configuration contiennent de moins en moins de documents. Un certain nombre de configurations n'ont pas d'histogrammes associés car aucun document commun n'a été trouvé pour permettre l'évaluation.

Ces valeurs mettent en perspective les performances indiquées, permettant par exemple de se rendre compte que si la configuration `0.1_1.4` peut sembler au premier abord sensiblement moins bonne que la configuration `0.3_1.4`, cela peut être en réalité une conséquence du fait que cette dernière est évaluée sur un ensemble trois fois plus petit. On réalise également que les configurations aux performances les plus élevées (à partir d'un seuil de similarité de 0,4) ne sont en réalité pas exploitables car leurs performances sont calculées sur un ensemble de données trop petit.

On observe par ailleurs que la pureté semble très sensible à cette variabilité, augmentant fortement lorsque le nombre de documents considérés pour l'évaluation diminue, tandis que la pureté inverse reste plutôt stable, ce qui tend à indiquer que notre baseline est en effet capable de séparer les documents en clusters distincts mais d'une granularité changeante, ne correspondant pas nécessairement aux types identifiés en référence. Compte tenu de ces éléments, nous nous abstiendrons de sélectionner

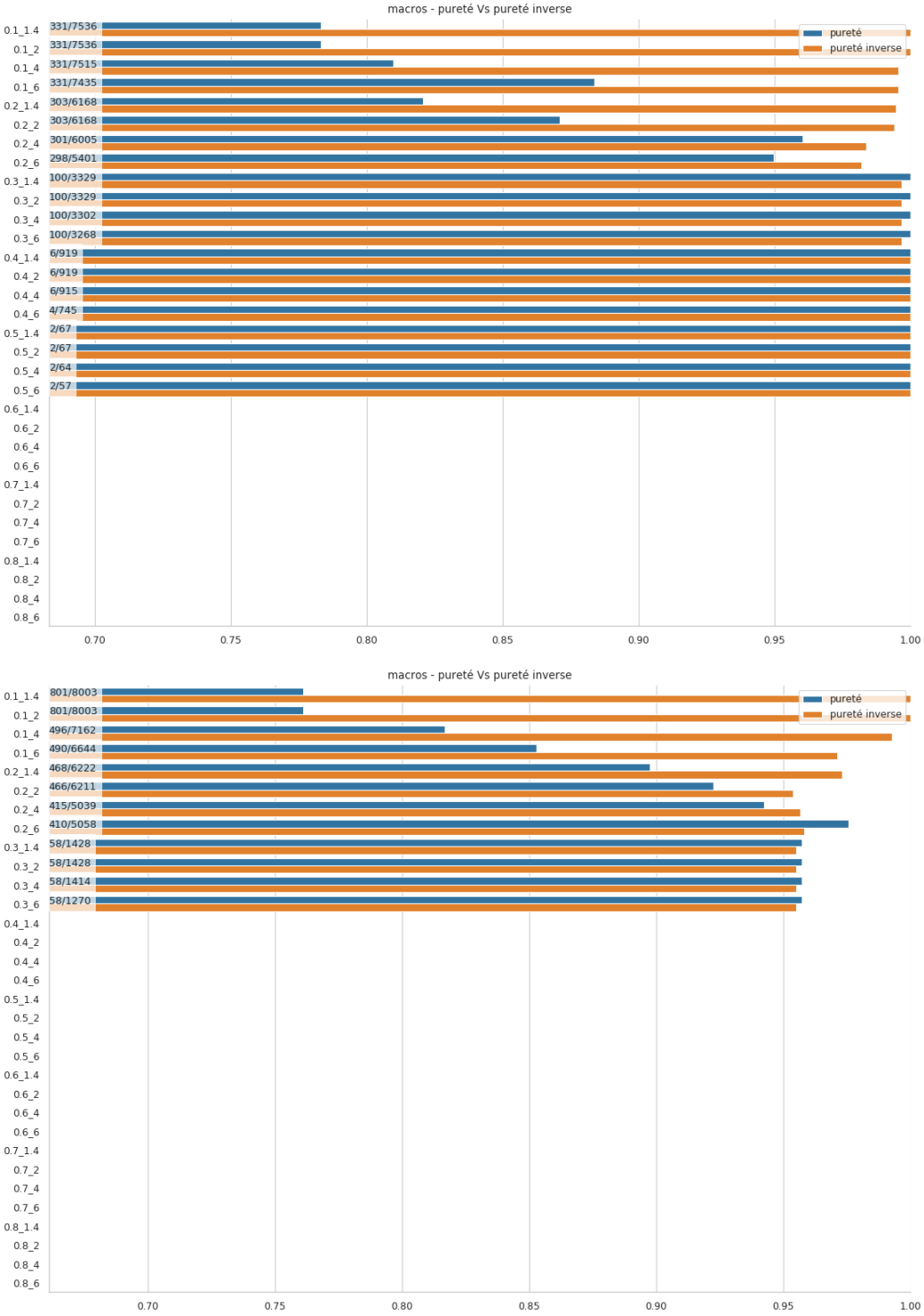


Fig. 4.5: Histogrammes des macro-puretés et macro-puretés inverses de notre baseline pour l'année 2015 et chaque configuration de paramètres, sur les Subject Codes comportant moins de 100 instances (en haut) et plus de 100 instances (en bas). Le seuil de similarité et le facteur d'inflation sont séparés par un underscore (ex : `0.1_1.4` indique un seuil de 0,1 et un facteur de 1,4). Les valeurs dans le cartouche blanc de chaque paire de barres d'histogramme indiquent le nombre de documents disponibles pour l'évaluation (recouvrement entre les documents dans nos résultats et la référence) / le nombre maximal de documents qui auraient pu être exploités (recouvrement maximal possible). Les configurations vides sont celles n'ayant pu être évaluées, faute de recouvrement.

une configuration baseline particulière pour la comparer aux performances de notre système, et préférerons une approche plus globale.

Les résultats de l'évaluation de notre méthode sont visibles en figure 4.6. Ici encore, nous ne montrons que les résultats pour la macro pureté, ceux de la micro pureté sont rapportés en annexe D. En ordonnée des graphiques, `sim:` décrit les modalités de calcul du score de similarité utilisées lors du clustering par MCL et `fusion:` indique celles de fusion des clusters contenant plus de 100 instances. La modalité `conditionnée` se rapporte au calcul original de la similarité entre mots d'instances différentes, dans lequel la composante syntaxique n'est calculée que si une similarité sémantique non nulle a été établie au préalable, tandis que `systématique` fait référence à notre variante consistant à calculer systématiquement ces deux composantes pour obtenir notre score final. Les modalités `recouvrement-lexical`, `recouvrement-fréquentiel` et `recouvrement-similarité` indiquent les performances des modalités de fusion des clusters telles qu'exposées en partie 4.3 et `aucune` correspond à ce nous appelons l'approche "sans fusion" (4.4.3), servant de baseline. Les Subject Codes n'ayant qu'un échantillon ne nécessitant pas de fusion de leurs clusters, ils dépendent seulement des modalités de calcul de la similarité. C'est pourquoi nous rapportons uniquement les deux performances associées.

On peut d'abord constater que le nombre de documents exploités pour l'évaluation est beaucoup plus élevé que pour la baseline, ce qui indique une meilleure adéquation entre les données couvertes dans nos résultats et celles de la référence. Le cas de la modalité `systématique` est particulièrement notable, produisant des ensembles pratiquement deux fois plus importants que la modalité `conditionnée` (dont les ensembles restent malgré tout beaucoup plus grands que notre baseline), ce que nous considérons comme une indication de la pertinence de notre stratégie consistant à relaxer les contraintes du calcul de similarité car les ensembles sont plus grands mais les performances se maintiennent à un niveau comparable. Ce plus grand volume de documents nous conforte par ailleurs plus globalement sur la pertinence de notre approche par rapport à la baseline étant donné que nos résultats restent bons et sont plus robustes, tant en termes de fiabilité (car ils portent sur plus de documents) que de variation entre les différentes configurations présentées.

À cette analyse quantitative s'ajoute le critère qualitatif non négligeable de l'exploitabilité des sorties. En effet, notre système produit des clusters de mots accompagnés d'informations de similarité lexicale très détaillées, constituant un matériau riche dans la perspective d'identifier des éléments caractéristiques des types d'événements induits susceptibles de constituer des schémas. À l'inverse, notre baseline ne produit que des clusters de documents nécessitant un travail complet d'extraction des éléments d'intérêt.

Pour terminer, nous préférons attendre de pouvoir mettre ces résultats en regard de ceux de l'évaluation des schémas que nous induirons à partir des types dégagés dans ces différentes configurations pour sélectionner une configuration particulière. Par conséquent, les performances de la méthode présentée dans ce chapitre sur les années 2013, 2014 et 2016 seront rapportées dans le prochain chapitre.

4.5 Conclusion

Dans cette phase, nous avons identifié des types d'événements à partir de descriptions d'instances. Pour ce faire, nous nous sommes détaché des éléments propres à la description d'instance au sens journalistique, à savoir une certaine redondance du discours et une composante temporelle marquée, pour exploiter des caractéristiques des types d'événements en adoptant une structuration thématique du corpus plutôt que temporelle et en explorant le relâchement de certaines contraintes sémantiques. Nous en tirons des clusters regroupant des ensembles de mots indépendants, structurés selon leurs caractéristiques syntaxiques et sémantiques et à même d'être utilisés dans la suite de nos travaux pour l'induction de schémas d'événements.

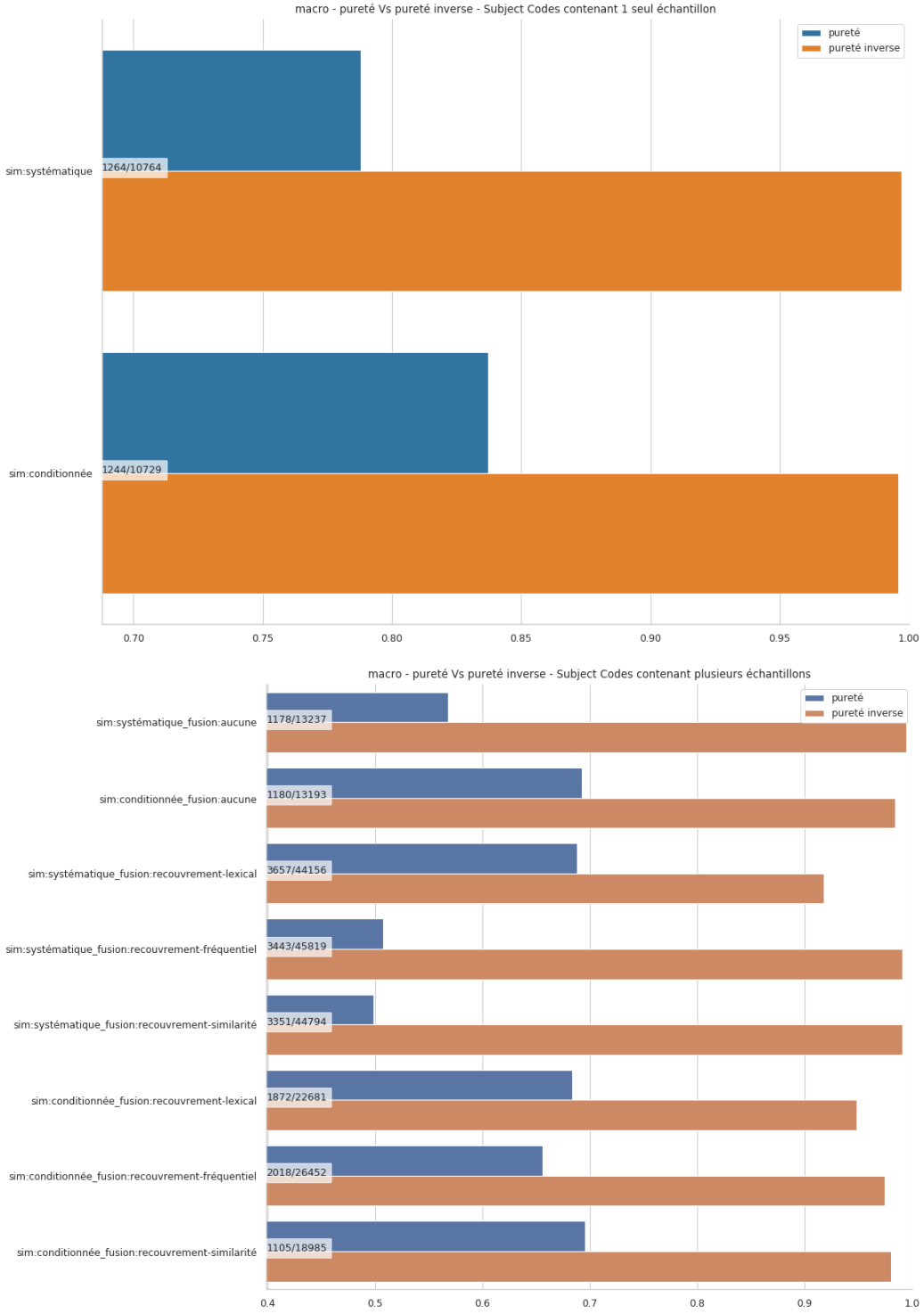


Fig. 4.6: Histogrammes des macro-puretés et macro-puretés inverses de notre méthode d'induction de types d'événements pour l'année 2015, sur les Subject Codes comportant moins de 100 instances (en haut) et plus de 100 instances (en bas). `sim:` décrit les modalités de calcul du score de similarité qui seront utilisées lors du clustering par MCL et `fusion:` indique les modalités de fusion des clusters contenant plus de 100 instances.

5 Du type au schéma

> (Perceval) - *Mais c'est quoi ça ?*
> (Hervé de Rinel) - *C'est l'île de Bretagne ! J'ai fait tout le tour deux fois pour être sûr !*
>
> **— Franck Pitiot et Tony Saba**
> Kaamelott, Livre IV, Les émancipés

Dans ce chapitre nous détaillerons la dernière partie de notre méthode qui vise, pour chaque type d'événement identifié précédemment, à extraire un schéma d'événement à partir des termes ayant émergé du processus de typage.
À l'issue de la première phase de traitement, nous disposons d'une description de chaque instance d'événement sous la forme d'une matrice d'alignement de termes telle que décrite en section 3.4. Dans la deuxième phase, nous lions ces matrices entre elles afin d'identifier des types d'événement. Nous calculons pour cela un score de similarité entre les lignes de chaque paire de matrices (et par extension d'instances) et nous utilisons la matrice de similarité résultante pour produire un clustering des instances en types d'événements. Ici, nous ré-exploitons les matrices de similarité entre les instances, ainsi que le détail des similarités entre les lignes des matrices qui les décrivent, pour construire des groupes de termes à l'échelle du type d'événement. Ces groupes forment la matière première à partir de laquelle nous structurerons nos schémas d'événement.

Notre méthode se fonde sur le parcours de graphes, une approche utilisée dans plusieurs travaux d'induction de schémas notamment (Nimishakavi et al., 2018), dans lequel un graphe triparti est parcouru pour unir différentes informations en un schéma cohérent, selon un principe comparable à celui qui motive notre propre approche. Le graphe est une représentation adoptée également dans (Kuzey et al., 2014), où chaque sommet correspond à un document et où les arêtes encodent à la fois une information d'ordre temporel et de similarité de contenu, il est ensuite transformé jusqu'à obtention des représentations visées.

L'approche que nous présentons se divise en trois étapes :

— l'extraction de proto-arguments sous la forme d'ensembles de termes à partir des types identifiés ;

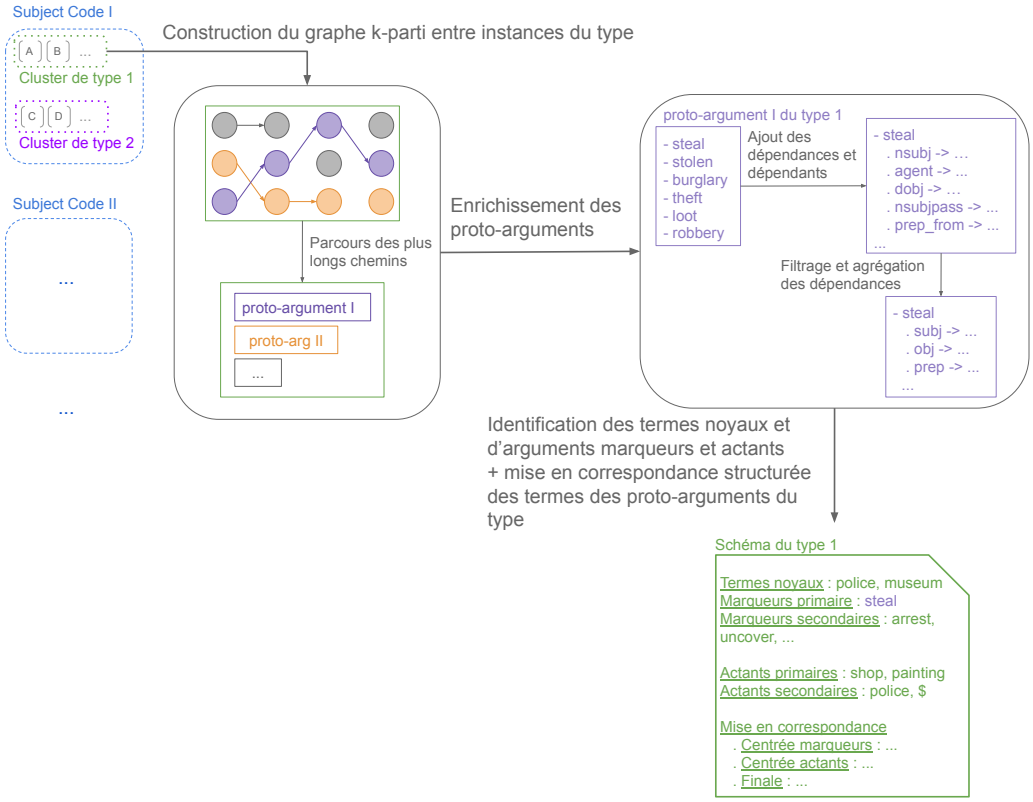


Fig. 5.1: Synthèse des traitements déployés pour induire un schéma d'événement à partir des informations définissant les types d'événements induits au chapitre précédent. L'étape impliquant la construction du graphe k-parti et son parcours est détaillée en section 5.1, l'enrichissement des proto-arguments en section 5.2 et la structuration des schémas en section 5.3

.

— l'enrichissement de ces proto-arguments avec des informations issues des contextes respectifs des termes qui les composent ;

— la structuration d'un schéma d'événement à partir de ces proto-arguments enrichis.

L'ensemble des traitements opérés dans ce chapitre sont résumés en figure 5.1.

5.1 Extraction de proto-arguments

L'objectif de cette étape est de dégager de grands ensembles de termes occupant la même fonction au sein du type d'événement considéré, c'est-à-dire grouper toutes les formes décrivant les victimes d'une catastrophe ou les objets d'un vol par exemple. L'idée sous-jacente est que chaque groupe ainsi identifié participe à la caractérisation du type de l'événement tout en pouvant être interprété indépendamment des autres, à la manière de chaque *slot* (argument) d'un *template* (schéma). Pour ce faire, nous construisons pour chaque type d'événement un graphe dans lequel chaque

sommet représente une ligne d'une matrice d'alignement décrivant une instance de ce type. Ces sommets sont reliés par des arêtes seulement si les deux lignes associées font partie de l'appariement final des deux instances (c'est-à-dire des appariements sélectionnés par l'algorithme hongrois). Ce graphe étant issu des résultats d'une succession d'appariements bipartis, il est par construction k-parti (où $k =$ nombre d'instances) et chaque sommet y est connecté avec un et un seul sommet de chaque autre partition. De plus, ses arêtes matérialisent des relations de similarité et peuvent être pondérées en utilisant le score associé, décrit en section 4.2. Un chemin de ce graphe va donc par construction relier des termes jouant des rôles similaires dans l'événement sous-jacent. On peut par conséquent considérer qu'un chemin dans ce graphe constitue un proto-argument potentiel pour la suite de nos travaux, et qu'il sera d'autant plus intéressant que le chemin en question est long (c'est-à-dire qu'il couvre plus d'instances).

Nous exploitons ces caractéristiques pour la construction de nos proto-arguments en commençant par appliquer récursivement un algorithme de calcul du plus long chemin dans le graphe, lequel supprime à chaque itération les arêtes constituant le chemin nouvellement trouvé. Ce processus est illustré dans la partie supérieure de la figure 5.2, qui représente le graphe : chaque ensemble de mots (que nous séparons par le caractère "|" car il s'agit des mots d'une ligne dans une matrice d'alignement) figure un nœud et les flèches constituent l'ensemble des arêtes. Le chemin le plus long de ce graphe est le chemin représenté en vert, qui relie les quatre instances représentée. L'identification de ce chemin entraîne la suppression dans le graphe de toutes les arêtes reliant l'un de ses nœud, c'est-à-dire celles reliant *steal*|*stolen* à *loot*|*steal* et *robbery*|*burglary* ainsi que *burglary*|*theft* à *robbery*|*burglary* et *steal*|*rob* à *loot*|*steal*. L'algorithme est réappliqué et trouve le chemin bleu et supprime l'arête entre *US*|*American* et *Italy*|*Italian*. Ce sont ensuite les chemins orange, jaune et rouge qui sont successivement identifiés (pas nécessairement dans cet ordre). On peut voir que cette approche simple permet de constituer un proto-argument autour de la dénomination de l'événement "vol" (en vert), un autre décrivant les entités qui réalisent cette action (en jaune), celles qui la subissent (en rouge) ainsi que la localisation de l'événement (en bleu). Pour appliquer cet algorithme, notre graphe doit être orienté et acyclique. Nous satisfaisons à ces contraintes en orientant les arêtes dans une même direction pour tout le graphe, le choix de la direction n'ayant pas d'incidence puisque chaque sommet est connecté à toutes les partitions.

À l'issue de cette étape, nous disposons de proto-arguments prenant la forme de groupes de termes remarquables vis-à-vis du type d'événement considéré. Cette opération de regroupement peut être envisagée comme une extension de l'identification de type en cela qu'elle repose directement sur les résultats de cette précédente étape, dans laquelle différentes instances d'événements sont appairées par le biais d'un score optimisant la similarité des termes qu'elles recouvrent. En assemblant

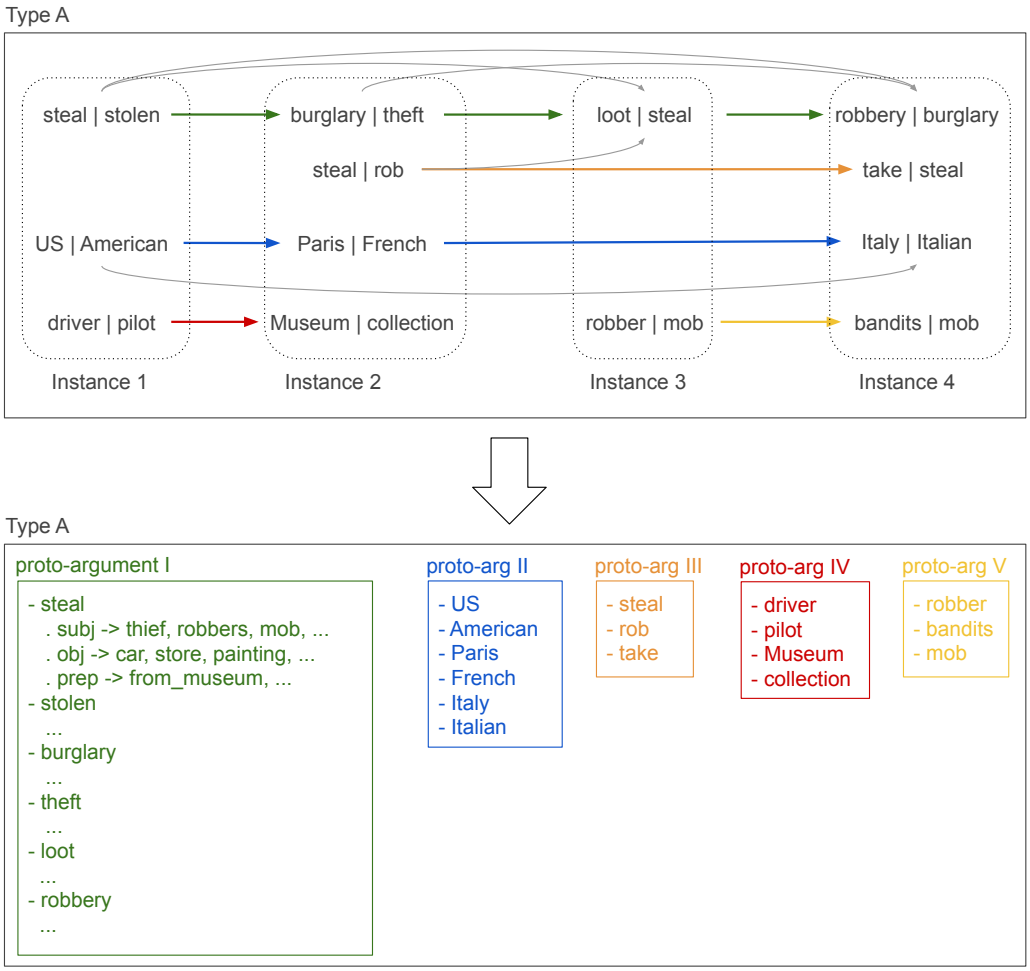


Fig. 5.2: Identification des proto-arguments. L'exemple est extrait de nos résultats pour le Subject Code "vol". Les flèches représentent le graphe k-parti des instances associées au type considéré. Celles en couleur indiquent les plus longs chemins trouvés, à partir desquels nous assemblons nos proto-arguments. Un exemple partiel des résultats de l'enrichissement des proto-arguments (détaillé en section 5.1) est donné pour le proto-argument I seulement, pour des raisons de lisibilité.

ces paires, nous construisons un graphe dont les propriétés nous permettent d'exploiter directement les informations de similarité sémantique et syntaxique déjà disponibles pour composer nos proto-arguments par identification récursive des plus longs chemins qu'il contient.

5.2 Enrichissement des proto-arguments par des informations contextuelles

Le regroupement de termes au sein d'un proto-argument repose sur un critère de similarité préexistant. Nous souhaitons amener plus d'informations dans chacune de ces structures afin d'améliorer la représentation qu'elles capturent de l'événement

Relations de dépendance syntaxique filtrées	Prépositions filtrées
dep, *aux*, *auxpass*, *cop*, *expl*, *goes-with*, *arg*, *poss*, *parataxis*, *punct*, *conj_and*, *conj_but*, *root*, *advmod*, *prt*, *det*	in, to, of, as, on, at, by, for

Tab. 5.1: Tableau récapitulatif des relations de dépendance syntaxique filtrées lors de l'enrichissement des proto-arguments.

et qu'elles constituent un socle riche dans la perspective de la construction de schémas d'événements. Pour ce faire, nous exploitons pour ce faire des informations propres aux contextes des termes regroupés, en ajoutant à chaque terme de chaque proto-argument ses dépendants syntaxiques dans les documents pertinents. Ainsi, chaque terme membre d'un proto-argument est rattaché à plusieurs ensembles d'autres termes répartis en fonction de leurs relations de dépendance syntaxique. Au cours de ce processus, nous conservons également l'information de la fréquence d'apparition de chaque terme du proto-argument dans le type. Celle des dépendants est conditionnée par la relation syntaxique qui leur est associée. Par ailleurs, nous filtrons certaines dépendances, dont la liste est indiquée dans le tableau 5.1, en raison de leur manque d'intérêt pour notre tâche. Enfin, nous souhaitons garder certaines prépositions (telles que *before* ou *after*) mais le jeu de relations originel impliquerait la création d'une entrée pour chaque type de préposition rencontrée, ce qui n'est pas utile. Nous les associons donc toutes à une entrée commune nommée `préposition`, dans laquelle nous stockons un terme combinant la préposition elle-même et le lemme du dépendant associé.
Cette étape supplémentaire permet notamment de récupérer des informations qui n'ont pas émergées en tant que proto-arguments individuels, comme le montre l'exemple de la figure 5.2, dans lequel on constate que le proto-argument I se voit complété par des informations relatives à l'objet du vol, exprimé par les mots *car*, *store*, *painting*, ..., ce qui enrichit la description de cet événement. Nous pensons également que les informations ainsi récupérées peuvent permettre d'établir des liens entre les proto-arguments, ce qui permettrait de donner une dimension plus holistique aux schémas induits.

Par ailleurs, afin d'améliorer la qualité de ces résultats, nous mettons en place deux heuristiques. La première consiste à filtrer les dépendances (et les dépendants associés) pour éliminer les moins intéressantes en ne sélectionnant que celles dont le terme de rattachement dans le proto-argument constitue plus de 80 % des occurrences. Le but de ce filtrage est d'éliminer les emplois marginaux d'un terme (tant sur l'axe syntagmatique que paradigmatique), qui pourraient sinon interférer avec

l'identification de sa fonction la plus courante (et donc la plus intéressante pour notre objectif).

Nous fusionnons ensuite les proto-arguments partageant au moins deux termes uniques et dont la somme des fréquences associées aux termes en commun est au moins égale à 50 % de la somme des fréquences des termes dans le proto-argument le plus gros (c'est-à-dire ayant le plus grand total d'occurrences entre les deux candidats). Ce traitement vise à remédier à certaines faiblesses de la stratégie d'induction de proto-arguments par calcul du plus long chemin dans un graphe, qui peut échouer à lier certains groupes de termes, essentiellement en raison d'appariements manqués lors de l'étape précédente.

Nous disposons alors de proto-arguments enrichis de leurs dépendants syntaxiques et de la nature de la dépendance qui les lient. Chaque terme d'un proto-argument est également décrit par sa fréquence dans le type d'événement considéré et chaque proto-argument, par le nombre d'instances qu'il recouvre. Dans l'étape suivante, nous tirerons parti de l'ensemble de ces informations dans le but de faire émerger, pour chaque type d'événement, une structure descriptive correspondant aux schémas d'événements qui constituent notre objectif final.

5.3 Structuration du schéma à partir des proto-arguments

Dans cette partie, nous exploitons l'ensemble des proto-arguments de chaque type d'événement, ainsi que les informations qui leur sont associées, dans le but de dégager une structure à même de remplir le rôle de schéma d'événement, c'est-à-dire de fournir une vision globale, synthétique et générique de l'événement considéré. Les proto-arguments sont de nature diverses : si l'on se réfère aux exemples de la figure 5.2, le proto-argument I pourrait être utilisé pour qualifier l'événement lui-même, alors que les proto-arguments IV et V en décrivent plutôt les participants. Nous souhaitons donc nous focaliser dans un premier temps sur l'identification, au sein des proto-arguments, de termes caractéristiques parmi trois catégories d'intérêt :

1. les termes *noyaux*, qui comme nous l'expliquons en partie 1.3, portent l'information du nom de l'événement. Le noyau d'un schéma induit sera constitué de deux termes au plus, qui devront être à la fois non ambigus et facilement interprétables ;
2. les termes *marqueurs*, qui identifient les actions et potentiels sous-événements constitutifs de l'événement considéré ;

Relations à gros grain	Relations originelles
acteur	*nsubj*, *agent*
objet	*dobj*, *iobj*, *pobj*, *nsubjpass*, *vmod*

Tab. 5.2: Tableau récapitulatif des relations de dépendance syntaxique originelles agrégées par notre jeu de relations à gros grain (rôles). Le jeu de relations originel est celui du *Stanford Parser*.

3. les termes *actants*, qui décrivent les participants impliqués dans ces actions et sous-événements.

Nous commencerons par catégoriser les proto-arguments dans leur entier, avant d'en extraire des termes représentatifs que nous considérerons comme les arguments de nos schémas. Afin de caractériser les proto-arguments, nous simplifions leur structure. Nous souhaitons en effet réduire le foisonnement de dépendances syntaxiques qui lient les termes d'un proto-argument à leurs dépendants tout en conservant suffisamment d'information pour produire des résultats intéressants. Dans cette optique, nous associons à chaque terme d'un proto-argument trois types de dépendances à gros grain [1], nommées "acteur", "objet" [2] et "préposition". Les rôles "acteur" et "objet" agrègent chacun plusieurs dépendances du jeu de relations originel, indiquées dans le tableau 5.2 (les relations non indiquées ainsi que les dépendants associés ne sont pas conservés dans les proto-arguments), tandis que "préposition" reprend directement le contenu de l'entrée commune du même nom construite lors de l'étape précédente. Cette double opération de filtrage et d'agrégation nous permet de centrer les proto-arguments autour d'une notion de rôle occupé dans l'événement.

5.3.1 Identification des termes noyaux

Comme nous l'avons évoqué en section précédente, nous envisageons l'identification des termes noyaux comme un moyen de nommer le type de l'événement associé au schéma induit. Dans cette perspective, nous pensons que le nombre de termes à rechercher doit être le plus restreint possible, et qu'ils doivent donc avoir une sémantique claire. Nous définissons d'abord les proto-arguments noyaux comme les deux proto-arguments couvrant le plus d'instances dans le type considéré. Les termes noyaux sont alors sélectionnés en retenant le terme le plus fréquent de chaque proto-argument noyau, sauf s'il s'agit d'un verbe ou d'un nom propre. Les verbes semblent des candidats intuitifs puisqu'ils sont les descripteurs naturels d'actions et par extension, d'événements, mais nous les excluons car ils font partie avec les noms communs des catégories les plus représentées dans les proto-arguments. Or,

1. Que nous appellerons dans la suite de ce document des rôles, bien qu'elles ne soit pas assimilables à des notions telles que celle de rôle sémantique.

2. Nous reprenons ici la terminologie utilisée dans (RADINSKY et al., 2012), la notion d'acteur se rapportant dans notre cas préférentiellement aux entités produisant les actions et les objets désignant le plus souvent celles sur lesquelles ces actions s'exercent.

nous souhaitons favoriser la sélection de noms communs car les dénominations les plus couramment attribuées aux types d'événements sont des formes nominales, plus précisément des nominalisations. Les noms propres sont quant à eux exclu en raison de leur caractère à la fois trop spécifique à certaines instances et trop général pour permettre de cerner les caractéristiques d'un type d'événement. En effet, il est courant (et normal) que différents types d'événements impliquent des entités de type `PERSON` ou `LOCATION`, lesquelles ne constituent que rarement, de fait, une information caractéristique d'un type particulier. Nous retenons deux termes plutôt qu'un car nous savons que tous les événements ne sont pas désignés par des nominalisations ou que ces dernières ne seront pas nécessairement assez fréquentes pour être sélectionnées. Dans ce cas, conserver deux termes offre un degré de liberté qui nous semble suffisant pour permettre de caractériser l'événement, sans nécessiter de chercher plus de termes ou de contraindre leur choix (par exemple en composant un verbe et l'un de ses arguments).

5.3.2 Identification des proto-arguments marqueurs

Nous considérons un proto-argument comme marqueur s'il décrit une action ou un sous-événement inscrit dans le type considéré. Cette notion se rapproche de celle de déclencheur lexical exploitée dans plusieurs travaux présentés plus tôt, elle-même issue des travaux sur les *frames*. Nous exploitons ce lien pour l'identification de nos marqueurs en nous appuyant sur la ressource FrameNet (Baker et al., 1998).

FrameNet est un projet initié en 1997 par Charles J. Fillmore et fondé sur sa théorie des cadres sémantiques (*semantic frames*). Un cadre sémantique est "une structure conceptuelle décrivant une situation, un objet ou un événement particulier, ainsi que les participants impliqués" [3]. Le projet FrameNet est une ressource décrivant 1 200 *frames* différentes pour la langue anglaise et a été adapté dans d'autres langues. Une *frame* comprend notamment un ensemble de termes appelés *lexical units*, regroupant des termes lemmatisés (ainsi que leur étiquette en partie du discours) dont le sens évoque la *frame* à laquelle ils sont associés.

Nous catégorisons un proto-argument comme marqueur si le terme le plus fréquent qui lui est associé est un verbe (indicateur naturel de l'action, de l'événement) ou si l'ensemble des *lexical units* d'une *frame* de FrameNet contient à la fois ce terme et un verbe. Cette dernière condition vise à permettre par exemple l'identification de nominalisations d'événements tout en évitant de prendre en compte certaines *frames* ne se rapportant pas à des situations ou des événements et qui ne sont donc pas pertinentes pour notre objectif. Nous avons en effet remarqué que les ensembles

3. "a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props" (Ruppenhofer et al., 2006).

de *lexical units* associés à ces *frames* non événementielles ne contenaient pas de verbes.

Par ailleurs, comme le nombre de proto-arguments marqueurs identifiés peut devenir important, nous souhaitons être en mesure d'en distinguer un sous-ensemble particulièrement saillant afin de produire des résultats faciles à exploiter, tant dans une logique d'interprétation humaine que d'évaluation automatisée. Pour ce faire, nous identifions le proto-argument marqueur dont le terme le plus fréquent est aussi le plus fréquent en valeur absolue. Puis nous cherchons parmi ses rôles (acteur et objet) les termes associés à d'autres proto-arguments marqueurs. Ce marqueur le plus fréquent associé à ceux qui se trouvent liés à lui forment ainsi les “marqueurs primaires", considérés comme les plus importants. Les autres proto-arguments marqueurs (non liés au plus fréquent) sont considérés comme des “marqueurs secondaires". Cette heuristique est fondée sur l'hypothèse que le terme marqueur le plus représenté est nécessairement central dans la description de l'événement, ce qui se traduit au niveau syntaxique par une position centrale dans la phrase autour de laquelle convergent les autres éléments caractéristiques.

Pour finir, le terme le plus fréquent de chaque proto-argument marqueur (primaire comme secondaire) est sélectionné comme terme représentatif et constitue un argument de notre schéma.

5.3.3 Identification des proto-arguments actants

Nous identifions les proto-arguments actants par opposition aux marqueurs. En effet, tout proto-argument qui n'est pas catégorisé comme un marqueur est considéré comme actant. Nous distinguons comme précédemment un sous-ensemble d'actants dits “primaires" dont l'identification s'appuie sur les marqueurs. Ainsi, tout proto-argument actant dont l'un des termes est représenté dans un rôle (acteur ou objet) d'un marqueur primaire est considéré comme un actant primaire, le reste des proto-arguments actants étant considérés comme actants secondaires. Ici encore, l'intuition qui motive cette heuristique repose sur le fait que les marqueurs primaires décrivent des actions ou des sous-événements notables et qu'il semble naturel d'appliquer cette mise en exergue à leurs participants. Les arguments actants sont retenus par sélection du terme le plus fréquent de chaque proto-argument actant, comme précédemment pour les autres proto-arguments.

La figure 5.3 résume visuellement ce processus sur un exemple. On peut voir que notre heuristique nous permet d'identifier *earthquake* comme un marqueur car il est inclus dans une *lexical unit* FrameNet contenant aussi un verbe (*shake*) mais d'exclure des noms communs tel que *threat* ou *coast* (du fait de l'absence de verbes

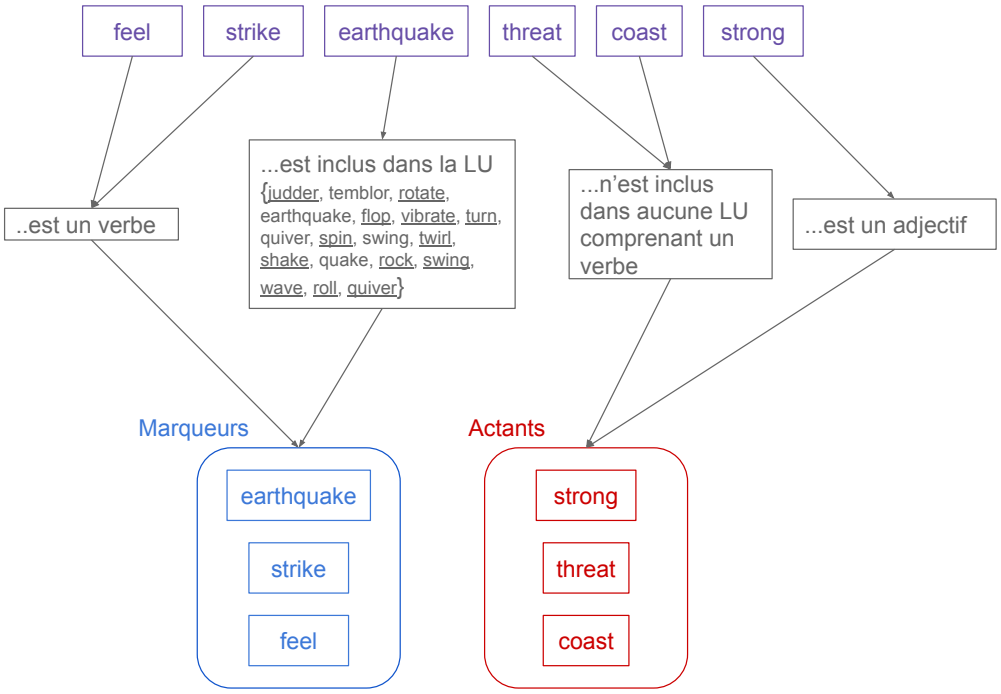


Fig. 5.3: Illustration du processus de catégorisation des termes représentatifs des proto-arguments en arguments marqueurs ou actants. Dans le détail de l'ensemble des LUs (*Lexical Units*) des *frames* FrameNet, les verbes sont soulignés.

dans les LUs qui leur sont associées), en faisant donc des actants. On remarquera que nous catégorisons les adjectifs comme des actants. Ce choix est fait car nous pensons que les adjectifs sont des termes intéressants à considérer mais qu'il serait inapproprié de les catégoriser comme marqueurs. De plus, considérant qu'ils sont plus souvent susceptibles de qualifier des actants, nous n'avons pas jugé nécessaire de leur dédier une catégorie.

5.3.4 Complément : correspondances entre actants et marqueurs

Comme nous l'avons expliqué dans les sections précédentes, nous considérons le terme le plus fréquent de chaque proto-argument actant et marqueur identifié comme autant d'arguments de nos schémas d'événements. Néanmoins, de par leur caractère induit, ces éléments ne sont pas équivalents aux arguments d'un schéma de type MUC et peuvent être difficiles à interpréter, que ce soit en raison de leur nombre ou de leur nature. Nous souhaitons donc compléter cette représentation par des structures à même d'expliciter les relations entre les marqueurs et les actants identifiés. Dans cette optique, nous mettons en correspondance marqueurs et actants dans trois structures constituées selon une méthode commune mais adoptant trois

perspectives distinctes. La figure 5.4 présente un exemple du schéma induit complet, avec ses termes noyaux, ses arguments et les différentes structures que nous nous apprêtons à détailler.

Chaque structure se compose de plusieurs entrées liant marqueurs et actants par le biais des relations (acteur ou objet) existantes dans les proto-arguments associés. Nous adoptons dans un premier temps une perspective centrée autour des proto-arguments marqueurs en faisant le lien avec les arguments actants trouvés parmi leurs dépendants et en précisant le type de la relation. La deuxième structure adopte quant à elle le point de vue des proto-arguments actants, identifiant de manière réciproque les arguments marqueurs trouvés parmi les dépendants ainsi que le rôle qui les lie. Dans ce cas, certains actants peuvent être des noms propres (souvent des noms de pays ou de personnes), qui sont alors normalisés par l'étiquette d'entité nommée correspondante.

L'induction de ces structures doit être effectuée séparément car elle repose sur le contenu des proto-arguments, qui ont été constitués indépendamment les uns des autres. De ce fait, si un actant A est présent dans les dépendants d'un proto-argument marqueur M, la réciproque n'est pas nécessairement vraie. Ce phénomène est visible au niveau de la figure 5.4, dans laquelle on peut voir que l'entrée liée à l'actant *police* est liée à trois marqueurs dans le premier cas et un seul dans le deuxième. Par ailleurs, il est possible qu'un dépendant (d'un marqueur ou d'un actant) apparaisse à la fois comme acteur et comme objet. Cela peut être dû à des erreurs d'analyse syntaxique, à des emplois marginaux du terme ou à des tournures de phrase ambiguës et se manifeste en particulier dans les types d'événements induits à partir d'un nombre restreint d'instances. Nous remédions à ce genre d'ambiguïtés en privilégiant la sélection du rôle où le dépendant est le plus fréquent.

Ces deux structures permettent d'appréhender clairement les liens entre les arguments identifiés mais ne rendent pas compte de la richesse lexicale disponible, non seulement parce que nous avons choisi de n'associer qu'un terme à chaque argument mais également parce que de nombreux termes présents dans les dépendants des proto-arguments pourraient apporter une information pertinente mais n'ont pas émergé comme proto-arguments indépendants dans les étapes antérieures. L'objectif de la dernière perspective adoptée consiste précisément à valoriser ces informations. Pour ce faire, nous souhaitons d'abord mentionner quelques constats empiriques réalisés lors de nos essais. Tout d'abord, le contenu des proto-arguments associés aux marqueurs nous a semblé apporter plus d'information sur la structure globale de l'événement. Toutefois, centrer la représentation des liens entre arguments autour des actants correspond mieux à nos objectifs. De plus, ces deux structures présentent beaucoup de redondance. Il n'est par conséquent satisfaisant ni de conserver les deux, ni de n'en retenir qu'une seule.

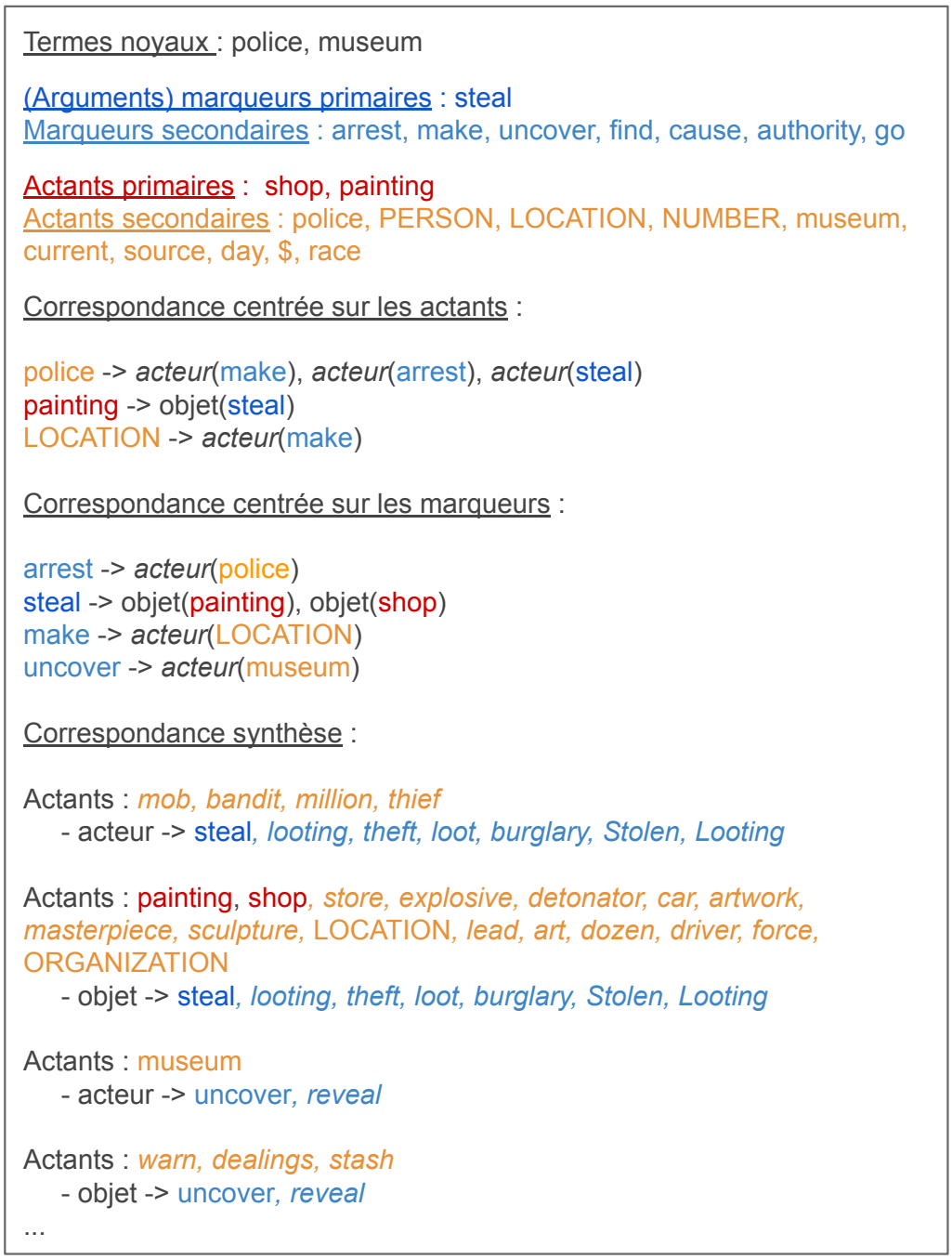


Fig. 5.4: Exemple d'un schéma induit, avec une représentation des différentes correspondances entre marqueurs et actants, dans le cas du Subject Code "vol". Dans les deux premières structures, le rôle acteur a été distingué du rôle objet par l'usage de l'italique. Dans la dernière structure, l'italique distingue les termes parmi les actants et les marqueurs qui ne font pas partis des ensembles identifiés dans les arguments.

Afin de faire la synthèse de ces deux perspectives complémentaires, notre dernière structure adopte une forme de compromis consistant à enrichir la structure centrée sur les marqueurs en y ajoutant les autres termes du proto-argument correspondant. Dans cette structure, les dépendants des proto-arguments associés sont essentiel-

lement des actants. Il est alors possible de centrer la perspective sur les actants en inversant la structure. La figure 5.4 présente également un exemple de cette structure. Dans cet exemple, les termes associés aux marqueurs et aux actants mais ne faisant pas partie des termes identifiés comme tels initialement ont été indiqués en italique afin de mettre en exergue la variété apportée par cette visualisation complémentaire. On peut ainsi voir que la deuxième entrée de cette structure (commençant par *painting*, *shop*...) est le miroir de la deuxième entrée de la correspondance centrée sur les marqueurs (dans ce cas, *steal*) mais présente une représentation notablement enrichie, à la fois dans ses actants et dans ses marqueurs. Cette inversion de point de vue permet donc d'associer des actants plus variés (s'agissant en fait des dépendants des marqueurs) et des marqueurs plus riches (complétés par leurs synonymes issus des proto-arguments). Le nombre d'entrées qui composent cette structure reste raisonnable puisqu'il sera au maximum de deux entrées par marqueurs (une par rôle, acteur ou objet), qui ne seront pas redondantes. En effet, chaque entrée sera unique, s'agissant d'un ensemble de dépendants, la redondance se situant au niveau des marqueurs associés. Cette redondance des marqueurs nous semble acceptable car elle est liée à l'ensemble des actions ou sous-événements dans lesquels les actants sont impliqués et il est naturel qu'une même action implique différents actants.

En conclusion, cette étape nous permet de structurer l'ensemble des proto-arguments à notre disposition pour chaque type d'événement afin d'en faire émerger un schéma. Pour ce faire, nous sélectionnons dans un premier temps deux termes particulièrement remarquables pour qualifier le type considéré. Puis nous simplifions le contenu des proto-arguments en ne conservant que les dépendances et dépendants syntaxiques associés aux termes du proto-argument par une relation d'acteur ou d'objet. Nous les répartissons ensuite en deux catégories, “marqueurs" et “actants", que nous avons raffinées en distinguant pour chacune d'elles des proto-arguments “primaires" et “secondaires". Nous considérons dès lors que le terme le plus fréquent associé à chaque proto-argument et caractérisé par sa catégorie (marqueur ou actant, primaire ou secondaire) représente un argument de notre schéma. Nous complétons cette description volontairement succincte par plusieurs déclinaisons d'une structure exploitant les liens entre proto-arguments pour expliciter les relations entre marqueurs et actants en exploitant au mieux la variété lexicale disponible dans les proto-arguments, selon plusieurs points de vue. Cette seconde partie n'a pas d'équivalent dans le format MUC dont nous nous inspirons mais permet une vision holistique du type d'événement qui complète (à notre avis) ce paradigme et nous permet de tirer un meilleur parti de nos résultats, considérant la variabilité inhérente à leur caractère induit.

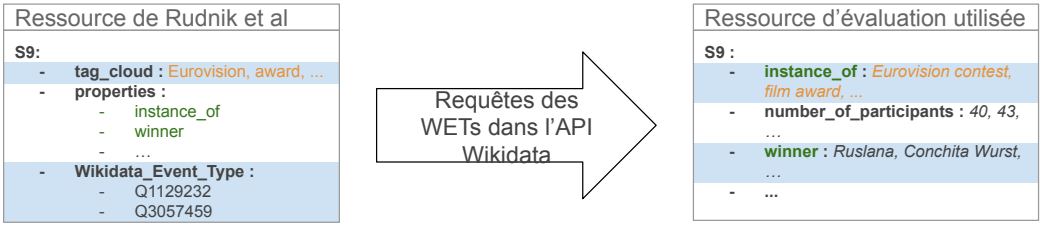


Fig. 5.5: Illustration du changement de format entre les schémas produits par RUDNIK et al. (2019) à gauche et notre représentation à droite. Le schéma de départ contient trois champs : `tag_cloud` (qui porte le contenu du champ `instance_of` des WETs agrégés), `properties` qui résume les noms (uniquement) des propriétés agrégées et `Wikidata_Event_Type`, qui porte les WETs agrégés. Notre représentation reconstruit à la fois l'ensemble des propriétés agrégées et leur contenu par le biais de l'API Wikidata.

5.4 Évaluation

5.4.1 Présentation de la ressource d'évaluation

Comme dans le reste de nos travaux, nous souhaitons évaluer les schémas d'événements induits par notre système selon des critères externes. Nous avons donc besoin d'une ressource de référence et faisons pour cela à nouveau usage des ressources de RUDNIK et al. (2019). En effet, le clustering final proposé par les auteurs regroupe différents événements à granularité fine de Wikidata en clusters de types plus généraux (par exemple un cluster regroupant différents sommets politiques). La description de chaque cluster est complétée par l'ensemble des propriétés Wikidata associées à chaque événement groupé, chacune pouvant être considérée comme un *slot*. Néanmoins, les noms de ces propriétés peuvent parfois s'avérer difficile à comparer au contenu de nos schémas (par exemple dans le cas des propriétés `located_in_the_administrative_territorial_entity` ou `armament`). De plus, les valeurs de ces propriétés, qui pourraient également être utiles, ne sont pas disponibles.

Afin d'obtenir une ressource d'évaluation plus adaptée à nos besoins, nous utilisons la liste des identifiants d'*items* comprise dans la description de chaque schéma pour effectuer des requêtes à l'API de Wikidata. Chaque requête retourne l'ensemble des propriétés décrivant l'*item* ainsi que les valeurs associées. Nous pouvons ainsi construire une ressource associant à chaque identifiant de schéma agrégé l'union des propriétés décrivant les types d'événements Wikidata liés à ce schéma ainsi que les valeurs associées (ou l'union des valeurs si une même propriété est renseignée dans plusieurs *items*). La figure 5.5 présente un exemple de ce changement de format.

5.4.2 Méthodologie

Nous souhaitons évaluer nos schémas d'événements par comparaison avec ceux de la ressource présentée en section précédente. Néanmoins, nos schémas se présentent sous la forme d'ensembles de termes associés à une fonction (marqueur ou actant) et de mises en correspondance entre ces termes et des termes similaires apportant une information de rôle (acteur ou objet). Il s'agit d'un format moins structuré que la collection de paires clé-valeur qui caractérise un schéma d'événement issu de Wikidata. Nous adaptons donc nos modalités d'évaluation afin d'obtenir des résultats permettant des analyses pertinentes de la qualité de nos schémas et exploitant au mieux les informations de chaque source.

Pour ce faire, nous identifions deux ensembles distincts dans ces deux types de schémas, que nous traitons séparément et utilisons de manière différente dans notre évaluation. Ainsi, pour chacun de nos schémas induits, nous comparons dans un premier temps les termes noyaux identifiés aux valeurs contenues dans le champ `instance_of` de chaque schéma de référence. Comme nous l'avons décrit en section 4.4.1, les valeurs de ce champ se rapportent à des dénominations d'événements (au niveau du type ou de l'instance) dans Wikidata. Or, c'est avec ce même objectif que nous avons repéré les termes noyaux dans nos propres schémas. Le principe de cette première étape est donc d'agir comme un filtre visant à circonscrire les traitements ultérieurs aux seuls schémas induits et de référence partageant au moins l'un des termes spécifiques à la dénomination de l'événement sous-jacent. La seconde étape consiste à mesurer le recouvrement entre les termes de notre troisième forme de mise en correspondance entre marqueurs et actants [4] et l'ensemble des termes constitué par les valeurs des *slots* du schéma de référence. Ces explications sont illustrées par un exemple en figure 5.6.

Précisons que les termes que nous manipulons du côté des schémas Wikidata sont des formes fléchies, les *slots* et leurs valeurs étant en général exprimées sous la forme de syntagmes courts (par exemple `number_of_injured` ou `magnitude_on_the_Richter_scale`). En revanche, nos schémas induits se composent de lemmes isolés. Pour pouvoir comparer ces deux formes, nous procédons avant toute chose à la racinisation de l'intégralité des termes utilisés.

4. C'est-à-dire centrée sur des ensembles de termes décrivant des actants, liés à des ensembles de termes désignant des marqueurs et associés l'un à l'autre par un rôle, acteur ou objet.

-instance of:	natural disaster, earthquake in New Zealand, Kamchatka earthquakes, earthquake
-maximum sustained winds:	115, 150, 130
-HURDAT identifier:	AL152010, AL052015
-topic's main category:	2010 Chile earthquake, Category:2011 Tohoku earthquake and tsunami, 2015 Nepal earthquake
-has effect:	Tsunami, Fukushima nuclear accident
-aftershocks:	1, 329, 1741
-earthquake magnitude on the Richter magnitude scale:	7.2, 8.3, 6.9
-duration:	15, 4
-JMA Magnitude:	7.3, 8.4
-located in time zone:	UTC-6:00
-damaged:	Church of Sant'Agostino, Fukushima Daini Nuclear Power Plant
-Commons maps category:	Maps of 2015 Nepal earthquake, Maps of 2016 Central Italy earthquake
-Wolfram Language entity code:	Entity[HistoricalEvent, April2015NepalEarthquake], Entity[HistoricalEvent, 2010HaitiEarthquake2010]
-part of the series:	Pacific hurricane, Atlantic hurricane
-number of injured:	4152, 293, 250000
-elevation above sea level:	-4, -11.4
-earthquake magnitude on the moment magnitude scale:	7.15, 8.4, 9.1
-cost of damage:	68680000000, 21500000000000, 108000000000
-destroyed:	building
-vertical depth:	212.5, 1.6, 19.7
-location:	Maule, Ryukyu Islands, Assiniboine River
-located on terrain feature:	Atlantic Ocean, North-East Pacific Ocean
-followed by:	2011 Tohoku earthquake and tsunami
-lowest atmospheric pressure:	995, 940, 998
-language of work or name:	English, Japanese, French
-described by source:	Q18024077
-number of missing:	2569, 0, 118
-number of survivors:	11
-JMA Seismic Intensity Scale:	shindo 7
-has cause:	rain
-permanent duplicated item:	Q19942159
-located in the administrative territorial entity:	Kamchatka Krai, Puerto Rico, Haida Gwaii
-part of:	2012 Atlantic hurricane season, 2016 & 2017 Central Italy earthquakes, 2005 Pacific hurricane season
-Commons category:	2017 Chiapas earthquake, 2016 Kaohsiung earthquake, 2015 Nepal earthquake
-on focus list of Wikimedia project:	WikiProject Zika Corpus
-subclass of:	recurrent event edition
-country:	Japan, South Korea, Nicaragua
-start time:	2016-06-23T00:00:00Z, 2016-04-14T00:00:00Z, 2012-08-11T00:00:00Z
-number of deaths:	217, 15894, 15896
-point in time:	2010-11-03T00:00:00Z, 2007-06-13T00:00:00Z, 2012-08-11T00:00:00Z
-end time:	2012-05-29T00:00:00Z, 2016-06-24T00:00:00Z, 2005-05-21T00:00:00Z
-Commons gallery:	2005 Kashmir earthquake, 2010 Baja California earthquake, Hurricane Katrina
-has immediate cause:	January 2017 Central Italy earthquakes, 2017 Farindola avalanche
-Mercalli intensity scale:	Mercalli scale VII, Mercalli scale X
-fault:	Nazca Plate, South American Plate

Termes noyaux : earthquake, people

Marqueurs primaires : earthquake, kill, end, disaster, move
Marqueurs secondaires : find, sever, chief, operation, urge, block, lose, pledge, pick, mark, possibility, battle, tourist, wait, authority smuggling, life, prayer, election, reopen

Actants primaires : people, LOCATION, aid, police, NUMBER, thousand, survivor, million, flood, one, community, year, landslide, Everest
Actants secondaires : Mount, massive, devastating, week, least, last, reconstruction, LOCATION, rubble, political, capital, NUMBER, child, $, ORGANIZATION, broken, poor, month, supplies, school, bath, northwestern, himalayan

Mise en correspondance Actant -> Marqueurs :

Actant : say, *thousand*, *landslide*, *LOCATION*, rise, disaster, avalanche, hundred, shake, battle, pull, face, check, avalanche, crash, *ORGANIZATION*, honor, slim, force, fading, strand, stop, bomb, forge, system, instruct, *one*

- acteur -> **kill**, survivor, death, dead, Secretary, claim, President, victim, party, capital, die, president, deadly, minister, leader, ministry, Leader, Victims

Actant : say, worker, rise, set, disaster, avalanche, trigger, *police*, shake, donor, toll, move, cause, *flood*, be, effort, pull, move, worker, official, bury, shift, fall, bring, team, struggle, race, flow, rip, *PERSON*, torrent, work, help, view, strand, staff, dozen, spur, stricken, wipe, live, response, four, topple, spark, reduce, cast, level, displace, flow, pose, plan, inflict, shatter, donation

- acteur -> **earthquake**, quake, hit, reach, quake-hit, devastate, Earthquake, **find**, earthquake-hit, destroy, devastating, ravage, damage, strike, come, damaged, sweep, devastation, injure, get, aftershock, Quake-Hit, Post-Quake, arrive, Earthquake-Hit, search, look, affect, seek

...

Fig. 5.6: Exemple des ensembles de termes comparés pour l'évaluation de nos schémas. En haut le schéma de référence, en bas le schéma induit candidat. En orange, les termes considérés comme "noyaux" de leurs schémas respectifs. La présence de termes noyaux communs conditionne le calcul du recouvrement entre les termes contenus dans les encadrés verts, qui déterminera le score de compatibilité entre les deux schémas. La comparaison est effectuée après racinisation de tous les termes.

Nous souhaitons produire une mesure quantitative de la compatibilité entre un schéma induit et un schéma de référence. Nous proposons pour cela une mesure combinant 3 facettes de cette notion de recouvrement, à savoir :

— le rapport entre le nombre de termes communs et le nombre de termes disponibles dans la mise en correspondance. Ce rapport nous indique la qualité de recouvrement du point de vue du schéma induit. Nous l'appelons $induced\_filler\_ratio$;

— le rapport entre le nombre de termes communs et le nombre de termes disponibles dans le schéma de référence. Il s'agit de la réciproque du rapport précédent, indiquant la qualité du recouvrement du point de vue du schéma de référence. Nous l'appelons $reference\_filler\_ratio$;

— le rapport entre le nombre de *slots* uniques mobilisés dans ce recouvrement et le nombre total de *slots* du schéma de référence. Cette proportion nous permettra de juger de la couverture globale du schéma, indépendamment du nombre de termes mis en jeu, qui peut fortement varier. Nous l'appelons $slot\_ratio$.

Notre score de compatibilité entre un schéma induit et un schéma de référence est alors la moyenne harmonique entre le recouvrement du point de vue des termes formant les valeurs des *slots* et la proportion de *slots* représentés :

$$compatibility(ind_i, ref_j) = \frac{2 * fillers\_mean * slot\_ratio}{fillers\_mean + slot\_ratio}$$

où $fillers\_mean = \frac{induced\_filler\_ratio + reference\_filler\_ratio}{2}$. Cette moyenne a été choisie pour sa tendance à défavoriser les écarts entre les éléments qu'elle combine. En effet, il nous semble important, à nombre de termes commun égal, de pouvoir valoriser un schéma induit couvrant plus de *slots*.

Ce calcul nous permet de produire, pour chaque schéma induit, une liste de schémas de référence triée par ordre décroissant de score de compatibilité. Nous choisissons donc de calculer l'*average-precision* pour chaque type induit, que nous agrégeons par le calcul de la *Mean Average Precision* (MAP) afin d'obtenir une mesure de la performance de chacune des huit configurations expérimentales décrites au chapitre précédent pour l'induction de types d'événements (en section 4.4.4 et visibles en figure 4.6). La MAP fournit une mesure globale de la qualité de l'ordonnancement produit par le système, que nous affinons par le calcul de la précision moyenne aux rangs 1, 2 et 3 pour chaque configuration. Enfin, comme le système peut produire une liste de schémas compatibles plus grande que le nombre de schémas annotés disponibles, nous calculons également la R-précision moyenne de chaque configuration, afin de compléter les résultats de la précision à différents rangs. Toutes ces mesures requièrent une référence annotée. Nous procédons donc à l'étiquetage de

certains schémas de référence en leur attribuant un Subject Code (information qui caractérise également chacun de nos schémas induits). Nous mettons en correspondance 13 Subject Codes différents avec 26 schémas de référence. À l'occasion de cette tâche, nous avons pu constater que certains schémas de référence étaient difficiles à exploiter. En effet, ces structures étant le résultat de l'application d'un algorithme de clustering, certaines d'entre elles regroupent des types d'événements parfois très différents et accumulent donc plusieurs collections de *slots* décrivant des événements totalement différents. Ce phénomène est illustré par la figure 5.7. Nous avons préféré ne pas utiliser ces schémas dans notre annotation, c'est-à-dire de ne les associer ni à un Subject Code ni à l'ensemble des Subject Codes auxquels ils auraient pu être rattachés.

Fig. 5.7: Exemple d'un schéma de référence agglomérant un grand nombre de types d'événements différents. Chaque type est identifié dans le champ `instance_of` et engendre donc l'agrégation d'une collection de *slots* associés, contribuant à l'émergence d'un schéma composite décrivant correctement chaque type, mais inexploitable du fait de sa diversité.

Les 13 Subject Codes retenus sont les suivants : "élection", "tremblement de terre", "catastrophes naturelles", "inondation", "vol", "incendie volontaire", "émeutes", "mouvements sociaux", "manifestation", "actes terroristes", "attaques à la bombe", "accident aérien" et "accident de train"[5]. Nous les avons choisis à la fois en fonction de l'existence d'au moins un schéma de référence associé mais également en essayant

5. D'après les noms originaux "election", "earthquake", "natural disasters", "flood", "theft", "arson", "riots", "civil unrest", "demonstration", "act of terror", "bombings", "air and space accident" et "railway accident".

de couvrir des types d'événements variés afin d'illustrer la capacité de notre système à représenter des événements différents, ce que les corpus habituellement utilisés, typiquement MUC, ne permettent pas du fait de leur trop grande homogénéité. Nous essayons également de mélanger des Subject Codes assimilables à des types d'événements (comme “tremblement de terre") et d'autres plus thématiques (comme “mouvements sociaux").

5.4.3 Résultats et discussion

La figure 5.8 indique les résultats de la MAP, la figure 5.9 les moyennes de la précision aux rangs 1, 2 et 3 et la figure 5.10 celles de la R-précision, le tout sur l'ensemble des Subject Codes évalués dans chaque configuration. On remarque que la précision à 1 (prec@1) est systématiquement la plus élevée. Cela indique que notre système associe la plupart du temps le type induit au sein d'un Subject Code au bon schéma de référence (dans notre annotation) avec la meilleure compatibilité. Les autres mesures de précision décroissent car il y a rarement plus de deux schémas de référence associés à un Subject Code. Les autres schémas compatibles sont donc considérés comme des erreurs, une tendance confirmée par le comportement de la R-précision.

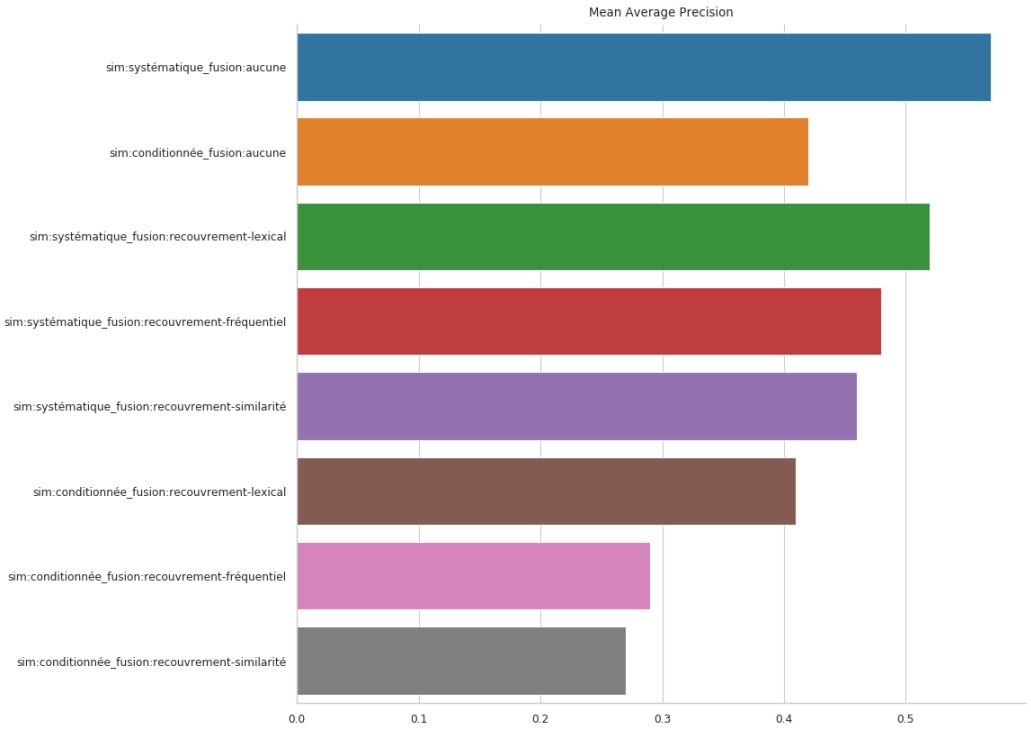


Fig. 5.8: Résultats de la *Mean Average Precision* (MAP) sur l'ensemble des configurations présentées pour l'étape d'induction des types d'événements.

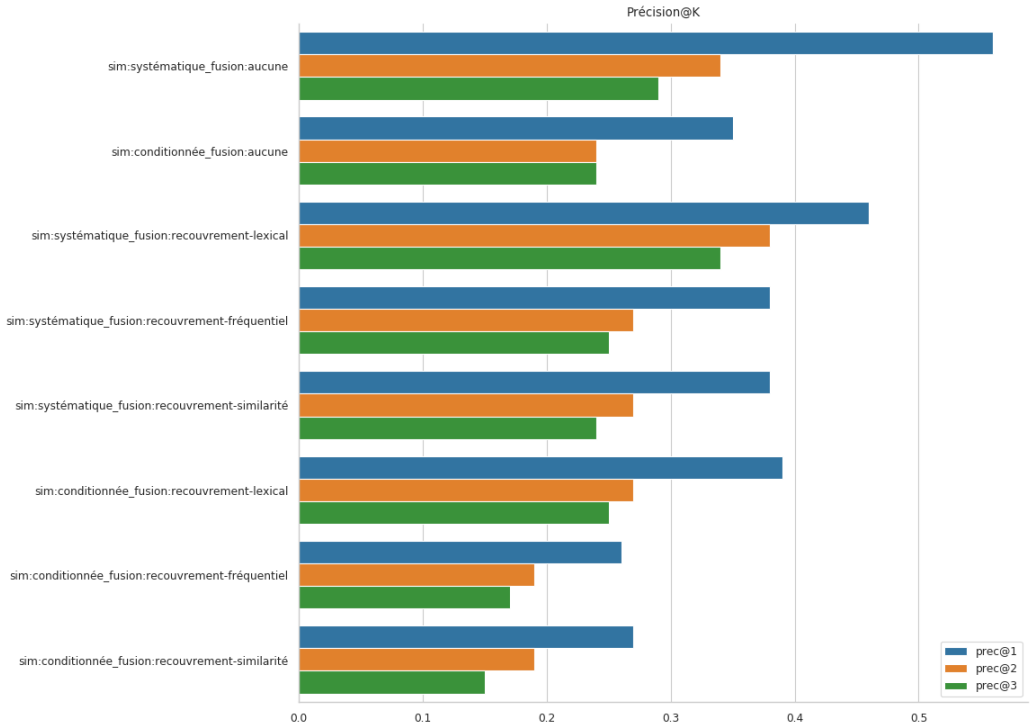


Fig. 5.9: Résultats de la Précision aux rangs 1, 2 et 3 sur l'ensemble des configurations présentées pour l'étape d'induction des types d'événements. Chaque valeur représente la moyenne des précision@k pour la configuration considérée. Exemple : prec@1 de `sim:systématique_fusion:aucune` est la moyenne des prec@1 de chaque Subject Code évalué dans `sim:systématique_fusion:aucune`

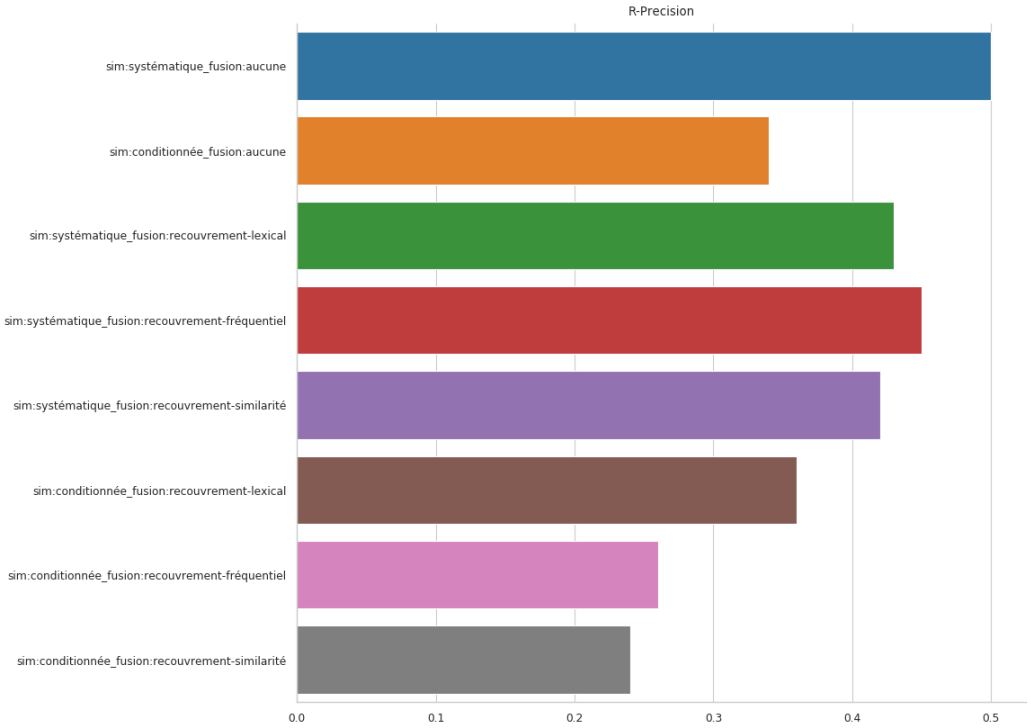


Fig. 5.10: Résultats de la R-Précision moyenne sur l'ensemble des configurations présentées pour l'étape d'induction des types d'événements. Chaque valeur représente la moyenne des R-Précisions pour la configuration considérée.

Deux configurations se détachent des autres en termes de performances, tant en ce qui concerne la MAP que les différentes mesures de précision : `sim:systématique_fusion:aucune` et `sim:systématique_fusion:recouvrement-lexical`. Commençons par souligner que ces deux configurations reposent sur notre stratégie de calcul systématique des composantes sémantique et syntaxique du score de similarité, ce qui constitue une indication supplémentaire (avec les résultats du chapitre précédent) de l'utilité de cette option. Si l'on examine la performance de ces configurations en termes de pureté (cf. figure 4.6 au chapitre précédent), on peut voir que `sim:systématique_fusion:aucune` ne conserve pas son avantage, avec une valeur absolue et un nombre de documents utilisés inférieurs. À l'inverse, `sim:systématique_fusion:recouvrement-lexical` fait partie des meilleures configurations selon ces deux critères et présente également l'avantage de mieux équilibrer pureté et pureté inverse. Toutefois la différence du nombre de documents à disposition pour l'évaluation de ces configurations (1 178 et 3 657 respectivement) appelle à la prudence. Nous proposons de combler le déficit d'information dû à cette différence de volumétrie de document par une comparaison plus qualitative des schémas obtenus dans ces deux cas pour deux Subject Codes de granularité différente, l'un pouvant être considéré comme un type d'événement et l'autre comme plus thématique. Nous choisissons le Subject Code “accident aérien" pour le premier

cas et “application de la loi" pour le second. Les comparaisons sont présentées en figures 5.11 et 5.12 respectivement.

Dans le Subject Code “accident aérien" (figure 5.11), on peut voir que la configuration `sim:systématique_fusion:recouvrement-lexical` produit un schéma plus compact, identifiant moins de marqueurs et d’actants secondaires et produisant moins d’entrées dans la mise en correspondance. Les concepts centraux sont néanmoins conservés : les marqueurs décrivent la survenue de l’accident et les recherches qui s’ensuivent, les actants impliqués sont l’engin accidenté, des lieux et des moments remarquables (*airport* et *landing*). La mention de l’entité `NUMBER` laisse à penser que le nombre de victimes est également capturé. Les deux configurations tissent des liens d’intérêt entre ces concepts, identifiant que ce sont par exemple des avions qui s’écrasent. Toutefois, la configuration `sim:systématique_fusion:recouvrement-lexical` se distingue en identifiant par exemple une relation entre l’accident et une perte (probablement d’ordre financier) pour la compagnie aérienne, en plus de l’accident. Les entrées montrent aussi une plus grande variété d’éléments caractéristiques, comme le lieu de l’accident (deuxième entrée, commençant par *runway*) ou les victimes (quatrième entrée), que ne montre pas la première configuration. Si ces éléments semblent jouer en faveur de la seconde configuration, il est néanmoins important de constater qu’elles donnent toutes les deux de bons résultats sur ce Subject Code “événementiel".

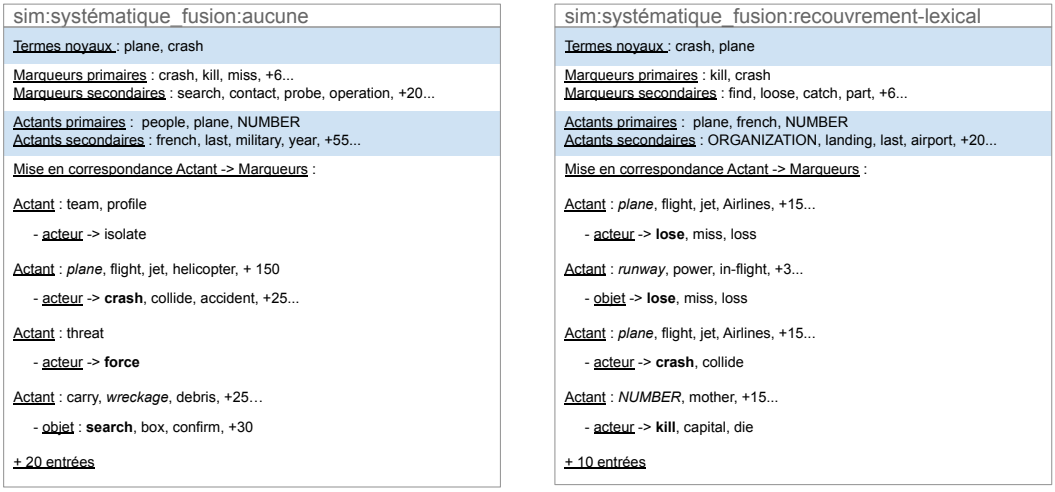


Fig. 5.11: Comparaison des schémas d’événements obtenus pour la catégorie “accident aérien" dans le cas de la configuration `sim:systématique_fusion:aucune` (à gauche) et `sim:systématique_fusion:recouvrement-lexical` (à droite). Cette dernière produit un résultat plus compact et de qualité équivalente en termes d’information décrite, ce qui en fait une configuration potentiellement plus intéressante.

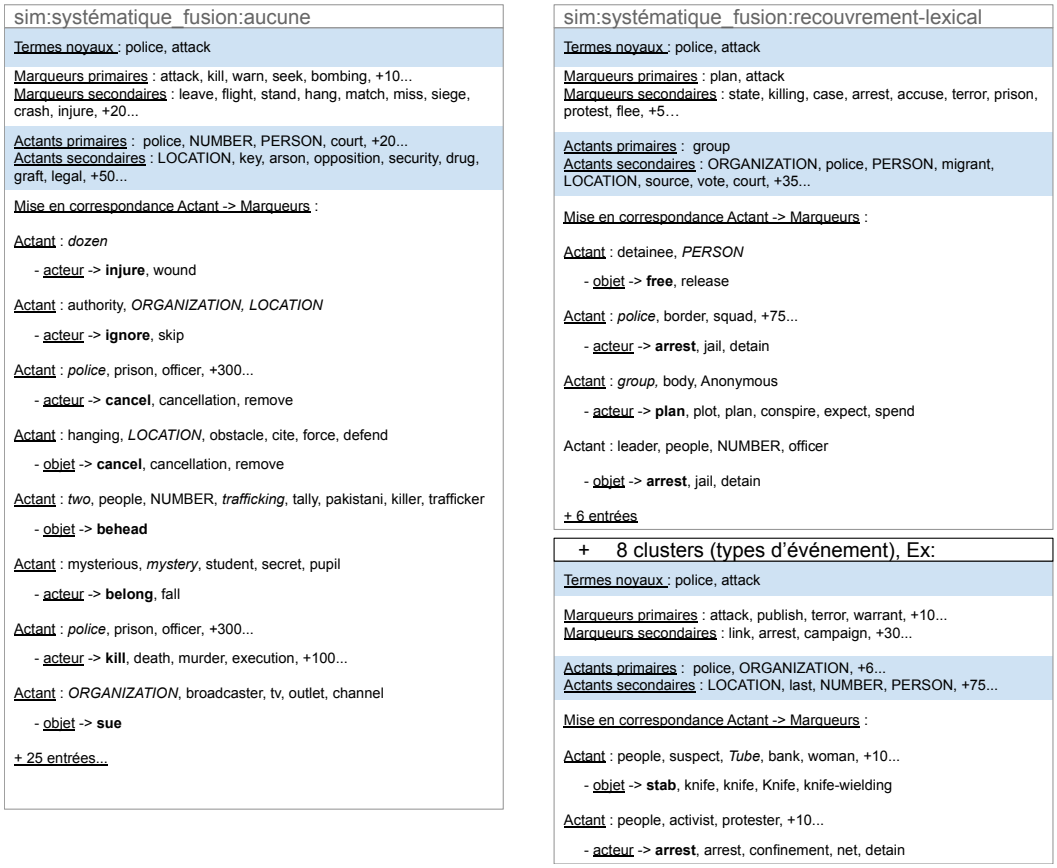


Fig. 5.12: Comparaison des schémas d'événements obtenus pour la catégorie "application de la loi" dans le cas de la configuration `sim:systématique_fusion:aucune` (à gauche) et `sim:systématique_fusion:recouvrement-lexical` (à droite). Cette dernière produit plus de clusters, dont deux sont visibles ici : le premier est présenté à titre de comparaison avec celui de gauche et le second pour illustrer le fait que ces différents clusters décrivent des types d'événements variés, alors que la première configuration ne produit qu'un seul cluster. Nous pensons que cette variété est un argument en faveur du choix de cette configuration.

Le Subject Code "application de la loi" (figure 5.12) est plus difficile à analyser aussi finement en raison du plus grand nombre de termes identifiés dans les deux configurations à n'importe quel niveau (plus de 50 actants secondaires dans `sim:systématique_fusion:aucune`, plus de 35 dans `sim:systématique_fusion:recouvrement-lexical...`). On constate en revanche que la configuration `sim:systématique_fusion:recouvrement-lexical` produit plus de clusters différents (neuf, contre un seul dans l'autre configuration), décrivant des types d'événements distincts (le premier cluster semble relever de la répression policière tandis que le second semble décrire des actions d'enquête et de protection en lien avec des attaques au couteau), ce qui correspond au comportement que nous attendons pour les Subject Codes thématiques tels que celui-ci. Par opposition, la configuration `sim:systématique_fusion:aucune` produit un seul cluster mélangeant meurtres, applications de la peine de mort et poursuites judiciaires.

À l'issue de ces analyses, nous proposons de retenir la configuration `sim:systématique_fusion:recouvrement-lexical`, qui nous semble présenter un meilleur compromis entre performances quantitatives et qualitatives. Nous appliquons notre système avec les paramètres associés sur le corpus de test des années 2013-2014-2016 et rapportons les performances obtenues au tableau 5.3, comparées à celles de 2015. On peut constater que l'ensemble des performances mesurées se maintiennent lors du passage sur cet ensemble de test, montrant même une légère amélioration, qu'il nous semble nécessaire de nuancer en raison de la baisse sensible du nombre de documents exploités pour l'évaluation des types d'événements. Néanmoins, ces résultats nous paraissent globalement d'autant plus satisfaisants que l'ensemble de test se compose de plusieurs années et présente donc intuitivement une plus grande variété événementielle que notre ensemble de validation (une seule année), ces résultats nous semblent donc indicatifs d'une bonne capacité générale de notre méthode à représenter des événements.

		2015	2013-2014-2016
Identification des types	Pureté	0,73	0,74
	Pureté inverse	0,95	0,97
	ratio de docs, cf. 4.4.4	4 921/54 920	2 882/43 895
Schémas	Mean Average Precision	0,52	0,58
	R-précision	0,43	0,46
	Précision @1/@2/@3	0,46/0,38/0,34	0,55/0,42/0,40

Tab. 5.3: Tableau des performances de nos méthode d'induction de types et de schémas pour la configuration retenue sur les années 2013-2014-2015, comparées à celles sur l'année 2015. Les valeurs de puretés rapportées sont une moyenne entre les valeurs macro, établies pour les Subject Codes de moins et de plus de 100 instances, présentés séparément dans les figures précédentes. Les ratios de documents sont obtenus en sommant les valeurs pour les deux modalités (plus et moins de 100 instances).

5.5 Conclusion

Dans ce chapitre, nous avons décrit la méthode par laquelle nous extrayons des informations caractéristiques des types d'événements que nous avons induits précédemment, et comment nous structurons ces informations dans une représentation synthétique, riche et interprétable, comparable à un schéma d'événement. Nous avons également présenté notre protocole d'évaluation de ces schémas, fondé sur la comparaison de leur contenu avec celui de schémas constitués par clustering à partir de données issues de Wikidata. Conscient des limitations d'une évaluation purement quantitative de notre travail, nous avons essayé de motiver nos choix par une analyse qualitative complémentaire. Ce travail d'évaluation présente toutefois l'avantage de porter sur une variété de types d'événements plus grande que les évaluations

MUC, qui constituaient la référence des corpus d'évaluation et dont nous tenions à dépasser l'homogénéité.

Conclusion et perspectives

6

> *Eh ben, on n'est pas sorti du sable !*
>
> **— Anne Girouard**
> Kaamelott, Livre IV, Les Novices

6.1 Synthèse des contributions

Dans ce travail, nous avons adopté une approche de Recherche d'Information (RI) dans le but d'induire des représentations synthétiques d'événements. Ces structures doivent permettent à des journalistes d'exploiter les grands volumes de données à leur disposition afin de produire de nouveaux contenus, dans la mouvance du journalisme de données. La méthode présentée réalise ce traitement sans entraînement préalable sur des exemples annotés, en combinant l'application d'algorithmes de clustering, de techniques de TAL et de parcours de graphes.

Notre approche repose sur une hiérarchisation des niveaux de description de l'événement qui conditionne les informations sur lesquelles nous focalisons notre attention à chaque étape. Dans un premier temps, nous nous concentrons sur l'identification de similarités lexicales et temporelles inhérentes à la description en continu d'un événement en train de se produire afin de regrouper les différentes mentions textuelles relatant la même instance, c'est-à-dire une réalisation particulière dans le temps et l'espace d'un événement. Nous nous détachons ensuite de ces ancrages forts au profit de la recherche d'équivalences sémantiques plus larges afin d'identifier les types d'événements en présence et de grouper les instances en fonction de ce nouveau critère. Enfin, nous extrayons et structurons les éléments les plus saillants de chaque type d'événement pour en produire une synthèse interprétable à la fois pour l'humain et pour la machine. Nous avons par ailleurs veillé à mener nos expérimentations sur d'importantes quantités de données, sans éliminer les cas extrêmes, tant pour produire les meilleurs résultats possibles que pour valider cette capacité de la méthode dans la perspective de sa réutilisation ultérieure.

Nos travaux s'inscrivent dans une direction de recherche explorée depuis longtemps, dont l'un des premiers éléments structurants a été la campagne MUC et ses structures

de référence appelées *templates*, se présentant sous la forme de collections de *slots* décrivant les concepts caractéristiques de l'événement considéré et constituant autant de clés auxquelles l'objectif était d'appairer les mentions textuelles s'y rapportant dans les textes. Bien que les structures que nous proposons ne présentent pas ce degré de lisibilité et de structuration, nous pensons qu'elles peuvent être exploitées dans des cas d'usages similaires. Dans la suite de ce chapitre, nous évoquerons quelques limitations de nos travaux avant de discuter d'un certain nombre de perspectives concernant les scénarios dans lesquels les résultats de ces travaux peuvent être mis en œuvre.

6.2 Limites des travaux présentés

L'une des limitations principales de notre approche réside dans son architecture en *pipeline*, où chaque module réutilise en entrée les résultats du module qui le précède. Ce couplage entraîne un phénomène de propagation et d'amplification des erreurs d'un module à l'autre, jusqu'au résultat final. Nous pensons que la mise en place de protocoles de validation de nos résultats à différentes étapes clés de la méthode a favorisé l'obtention de résultats finaux de meilleure qualité que si nous avions cherché à optimiser l'ensemble de nos modèles uniquement en nous fiant à ses performances sur l'évaluation des schémas, en particulier car nous disposons de plus de données d'évaluation aux niveaux intermédiaires qu'au niveau final. Nous pensons également que ces évaluations multiples ont contribué à contenir les effets délétères de la propagation d'erreurs tout au long du processus de développement, ce qui nous permet par exemple d'envisager nos descriptions d'instances comme des structures exploitables en elles-mêmes plutôt que comme un simple résultat intermédiaire. De plus, nous avons adopté cette architecture car nous pensons qu'elle présente malgré tout certains avantages par rapport aux architectures intégrées (*end-to-end*). En effet, la tâche d'induction de schémas d'événements est intrinsèquement complexe et nous pensons qu'une architecture en *pipeline* permet de confiner cette complexité dans les différents modules, ce qui permet *in fine* d'intégrer plus simplement et explicitement différents éléments de modélisation et de contrôler plus finement leurs effets sur le résultat final.

Une autre limite de notre méthode se situe au niveau des représentations manipulées. En effet, si nous tirons parti des volumes de données à notre disposition par l'utilisation d'un certain nombre de méthodes statistiques, nos traitements reposent sur une représentation symbolique des textes considérés. Ce choix s'explique en partie par le fait qu'il permet d'incorporer facilement des traitements linguistiques par le biais d'outils tels que les analyseurs syntaxiques. Ces outils deviennent néanmoins, par effet de bord, un facteur potentiellement limitant de notre méthode à la fois par

leurs performances mais aussi par les erreurs qu'ils introduisent et qui s'ajoutent à celles de nos propres traitements. Le recours à des représentations symboliques est également motivé par le manque de représentations denses adaptées à nos besoins au début de nos travaux.

En effet, les premiers plongements lexicaux à avoir connu un fort succès dans les dernières années étaient ceux issus de modèles tels que Word2Vec (MIKOLOV et al., 2013) ou GloVe (PENNINGTON et al., 2014) [1]. Par ailleurs, plusieurs études (LEVY et al., 2015 ; D. CHEN et al., 2017) ont par la suite apporté un certain nombre de nuances sur l'intérêt de ces représentations par rapport à des représentations de type sac de mots, ce qui nous a encouragé à privilégier dans nos propres travaux l'usage de représentations symboliques, qui présentaient le double avantage d'être mieux maîtrisées et de nous laisser plus de liberté dans les traitements auxquels nous pouvions recourir. Parallèlement, le domaine de l'apprentissage de plongements lexicaux s'est fortement développé, avec notamment l'introduction des plongements contextuels de type ELMo (PETERS et al., 2018) ou BERT (DEVLIN et al., 2018) et un apport significatif pour tout un ensemble de tâches dans le domaine du TAL. La question de l'intégration de ce type de représentations dans notre approche se pose donc légitimement et plus précisément, dans notre cas, pour la similarité de phrases au niveau du clustering en instances et de l'alignement monolingue de phrases pour le filtrage des dépêches, cet alignement conditionnant une part importante des traitements ultérieurs au travers des matrices d'alignement. Concernant la similarité des phrases, des travaux récents permettant d'évaluer cette similarité efficacement à grande échelle avec des plongements de type BERT (REIMERS et GUREVYCH, 2019) constituent une piste intéressante dont l'intérêt mériterait d'être évalué dans notre contexte. Les travaux traitant de l'alignement monolingue de phrases sont pour leur part nettement moins nombreux que ceux concernant la similarité de phrases mais des approches neuronales récentes (OUYANG et MCKEOWN, 2019) représentent une piste à suivre. Par ailleurs, compte tenu de la proximité de l'alignement monolingue de phrases avec les travaux sur la traduction automatique et le succès pour cette dernière des modèles *transformers* (VASWANI et al., 2017), également à l'œuvre dans le cas de BERT, il est raisonnable de penser que cette tâche pourrait bénéficier de l'introduction de ce type de modèles, progrès dont bénéficierait par extension notre approche.

1. Nous insistons ici sur le fait que nous sommes pleinement conscient qu'il ne s'agit pas des premières propositions dans le domaine historiquement mais des premières (à notre connaissance) à avoir suscité un intérêt massif au sein de la communauté du TAL depuis de nombreuses années.

6.3 Peuplement d'une base de connaissances

Les travaux présentés ont été réalisés dans le cadre du projet ASRAEL, en partenariat avec plusieurs acteurs académiques et industriels, dont le centre de recherche rattaché à l'école d'ingénieur EURECOM pour son expertise dans le domaine du Web sémantique. L'un des objectifs de cette collaboration est d'intégrer le résultat de nos travaux à une base de connaissances. Sa réalisation permettrait par exemple le peuplement automatique de cette base à partir de nos résultats, une tâche explorée notamment par la piste KBP (*Knowledge Base Population*) de la conférence TAC (Text Analysis Conference), à la différence que ces tâches reposent sur des corpus annotés qui définissent et limitent de fait les événements et les éléments d'informations considérés. Notre approche permettrait de ne plus dépendre de ces corpus et de lever du même coup ces limitations.

Le déploiement d'une base de connaissances intégrant ces schémas est par ailleurs une étape importante du processus de "dataification" des contenus journalistiques, lui-même un jalon majeur pour la mise en place de pratiques de journalisme des données, qui constitue l'une des motivations principales de l'implication de l'AFP dans le projet ASRAEL.

6.4 Moteur de recherche orienté événement

Un cas d'usage envisagé dès le début du projet afin de tirer parti d'une intégration de nos schémas dans une base de connaissances est celui d'un moteur de recherche centré autour de la notion d'événement. Dans ce scénario, l'utilisateur entrerait une requête qui lui retournerait dans un premier temps différents schémas pertinents. L'utilisateur sélectionnerait alors le type d'événement qui l'intéresse et pourrait ensuite indiquer quels éléments caractéristiques du schéma il souhaite intégrer à sa recherche. Cette nouvelle requête retournerait alors un sous-ensemble de documents ciblant non seulement un type d'événement particulier mais tenant également compte du besoin informationnel spécifique de l'utilisateur. L'avantage d'une telle interface est de combiner l'intuitivité des moteurs de recherche en langage naturel avec la structuration qu'apportent les systèmes du Web sémantique afin de permettre à des utilisateurs spécialisés de naviguer efficacement dans de grandes masses de données, ce qui constitue un autre enjeu majeur du journalisme de données. Les travaux de RUDNIK et al. (2019) présentent une version préliminaire de cet outil et l'intégration des résultats du présent travail est un objectif futur du projet ASRAEL.

6.5 “Schématisation" à la volée d’un nouveau document

Un autre cas d’usage envisageable de nos schémas consisterait à comparer toute nouvelle dépêche émise aux différents schémas induits afin de déterminer le plus représentatif. Il s’agirait donc d’un processus “en ligne" à l’issue duquel les différents éléments constitutifs du schéma (marqueurs, actants...) pourraient être utilisés pour mettre en exergue les mentions notables dans la dépêche. Ce travail d’aide à la visualisation a pour but de faciliter la tâche de journalistes ayant à synthétiser un grand nombre de dépêches. Pour ce faire, une approche fondée sur la correspondance de formes dans le schéma et dans la mention peut être envisagée comme un point de départ qui pourrait être approfondi par la suite par des méthodes exploitant des mesures de similarité. On peut, dans cette perspective, envisager les schémas induits comme la matière première pour la construction de représentations denses d’événements, une tâche déjà explorée par exemple par MODI (2016) dans le cas de l’induction de *scripts*.

Dans la perspective d’une amélioration continue du système, une telle interface pourrait également être exploitée pour permettre aux journalistes d’indiquer quelles sont les mentions réellement utiles du document (éliminant les faux positifs et signalant les faux négatifs) afin de faire évoluer le contenu de nos schémas pour les adapter aux besoins des utilisateurs.

6.6 Évaluation par comparaison avec d’autres systèmes

Nous nous sommes efforcé de mettre en place dans nos travaux des protocoles d’évaluation à même de nous fournir des informations sur la qualité de nos résultats en tenant compte à la fois de la quantité et de la variété des événements que nous traitons. Nous pensons également qu’en dépit de leurs limitations, ils nous ont permis de montrer que la méthode proposée conduit à l’obtention de schémas satisfaisants. Le dernier protocole présenté, visant à évaluer les schémas en tant que tels, reprend la forme de celui proposé par CHAMBERS et JURAFSKY (2011) en cela que nous nous appuyons sur la mise en correspondance des valeurs associées à différents *slots* de référence avec les éléments de nos structures candidates pour attribuer à ces dernières l’étiquette de référence maximisant un score de similarité. Nous l’avons choisi en partie parce que la définition d’une meilleure alternative est une tâche difficile. En effet, comme nous l’avons déjà discuté, le résultat d’un système de

traitement automatique des événements est fortement conditionné par la définition de la notion d'événement retenue par les concepteurs, ce qui rend difficile la mise en place d'un protocole unique et équitable dans une configuration non supervisée. De plus, ce protocole est le seul à avoir été repris par d'autres travaux. Il serait dès lors intéressant de comparer la performance des systèmes issus de ces différents travaux sur la base de données aussi hétérogènes que celles que nous avons manipulées afin de dépasser la limitation principale de ces évaluations, à savoir leur recours au corpus MUC, trop petit et trop homogène pour permettre d'estimer la performance des systèmes évalués dans les cas d'usage qu'ils ambitionnent. Cette limitation est par ailleurs la raison pour laquelle nous n'appliquons pas ce protocole directement à nos propres travaux. Nous n'avons malheureusement pas été en mesure de mener ce travail nous-même faute de temps, notamment en raison du coût que représente une telle entreprise, beaucoup des systèmes décrits ne disposant pas d'implémentations prêtes à l'emploi.

Table des figures

Liste des tableaux

A Guide d'annotation d'instances d'événements

A.1 Introduction

L'annotation de document comme appartenant à une instance d'un événement donné porte sur les documents issus du Web, dans le cadre de l'expérimentation sur le filtrage des clusters hybrides. Pour faciliter la constitution d'un corpus d'annotation, nous avons lancé le module une première fois (entrée : clusters hybrides) pour obtenir des clusters d'instances triés par score d'alignement (sortie). Nous avons sélectionné les 50 premiers documents (les mieux notés) et les avons annotés.

L'annotation d'un document Web se fait toujours en le comparant uniquement à la référence AFP du cluster auquel il a été attribué.

Il y a trois niveaux de similarité avec la référence, associés à trois valeurs à renseigner dans l'attribut `relevance` :

— 2 : Le document relate le même événement que la référence

— 1 : Le document est fortement lié à l'événement de référence mais ne traite pas complètement de l'événement

— 0 : Le document est complètement différent de la référence

A.2 Annotation "*relation forte*" (valeur 2)

Un document Web est considéré comme faisant partie de la même instance d'événement que la référence AFP si :

— Le sujet du document est centré sur le même événement, et le contexte spatio-temporel est le même ;

— Le titre et le chapeau du document sont cohérents, c'est-à-dire qu'ils remplissent individuellement la condition précédente, par exemple :

(A.1) a. Référence AFP :
Truck bomb kills at least 33 in Baghdad market.

A truck bomb ripped through a market in a Shiite-majority area of north Baghdad on Thursday, killing at least 33 people, security and medical officials said.

b. Document valide pour une annotation de valeur 2 :
Dozens Killed in Baghdad's Sadr City as Massive Truck Bomb Explodes.
Truck detonated in Sadr City's Jameela market, a predominantly Shiite neighborhood, shortly after dawn.

— Le titre et le chapeau ne sont pas cohérents, mais le titre identifie clairement l'événement de référence et le chapeau se situe dans le même contexte, par exemple :

(A.2) a. Référence AFP :
Wildfires rampage across northern California.
Firefighters on Monday battled devastating wildfires that have reduced hundreds of homes to smoldering ruins and threatened California's renowned wine region

b. Document valide pour une annotation de valeur 2 :
Hundreds of homes go up in flames in California wildfires.
The Napa County Fairgrounds is usually a place you go to have fun – to watch a race, enjoy a show or revel at a festival.

A.3 Annotation "*relation distante*" (valeur 1)

Un document est considéré comme lié à la référence AFP si :

— Le document relate le même événement que la référence (c'est-à-dire que le type et le contexte spatio-temporel sont les mêmes) ;

— Le titre et le chapeau ne sont pas cohérents car le titre est générique, mais l'événement de référence est clairement relaté dans le chapeau, par exemple :

(A.3) a. Référence AFP :
Wildfires rampage across northern California.
Firefighters on Monday battled devastating wildfires that have reduced hundreds of homes to smoldering ruins and threatened California's renowned wine region

b. Document valide pour une annotation de valeur 1 :
AP NewsAlert.
MIDDLETOWN, Calif. (AP) — State officials : 23 000 displaced from 2 massive wildfires sweeping Northern California.

— L'information apportée par le document n'aurait pas de sens sans la mention de l'événement de référence. Les trois stéréotypes de cette situation sont :

— Les bilans de victimes d'une catastrophe naturelle ou non (*death toll*). Sont concernées les articles en faisant mention et ceux mettant à jour les chiffres du bilan, par exemple :

(A.4) a. Référence AFP :
Magnitude 7,5 earthquake hits Nepal : USGS.
A powerful 7,5 magnitude earthquake struck Nepal on Saturday, the United States Geological Survey said, with strong tremors felt across the Himalayan nation and parts of India.

b. Document valide pour une annotation de valeur 1 :
Nepal earthquake toll rises to 1 341 : police.
NEW DELHI (Reuters) - The death toll in Nepal from a severe earthquake on Saturday has risen to 1 341, a police spokesman said.

— Le décret d'un statut administratif particulier, par exemple le passage en état d'urgence après une catastrophe :

(A.5) a. Référence AFP :
Wildfires rampage across northern California.
Firefighters on Monday battled devastating wildfires that have reduced hundreds of homes to smoldering ruins and threatened California's renowned wine region.

b. Document valide pour une annotation de valeur 1 :
California fires displace 23 000 people.
The governor of California has declared a state of emergency after wildfires forced about 23 000 people to flee their homes in the north of the state.

— La revendication d'une attaque par un groupe :

(A.6) a. Référence AFP :
Truck bomb kills at least 33 in Baghdad market. A truck bomb ripped through a market in a Shiite-majority area of north Baghdad on Thursday, killing at least 33 people, security and medical officials said.

b. Document valide pour une annotation de valeur 1 :
Islamic State claims truck car bomb in Baghdad's Sadr City : statement. CAIRO (Reuters) - Islamic State claimed responsibility on Thursday for a truck bomb attack at a crowded marketplace in the Baghdad district of Sadr City which killed at least 76 people and wounded more than 200 others.

Exemples complémentaires :

(A.7) a. Référence AFP :
Wildfires rampage across northern California.

Firefighters on Monday battled devastating wildfires that have reduced hundreds of homes to smoldering ruins and threatened California's renowned wine region.

b. Document valide pour une annotation de valeur 1 :
What It Was Like Inside California's Raging Valley Fire.
Raging fires spread throughout Middletown, California, Sunday, and while most of the residents had fled for their safety, the rescue crews were left looking for unexpected hazards as they fought the flames.

c. Document valide pour une annotation de valeur 1 :
10 Things to Know for Tuesday.
Power lines continue to burn along Highway 175 outside Middleton, Calif., Monday, Sept. 14, 2015, Two of California's fastest-burning wildfires in decades overtook several Northern California towns, killing at least one person and destroying hundreds of homes and businesses and sending thousands of residents fleeing highways lined with buildings, guardrails and cars still in flames.

A.4 Annotation "négative" (valeur 0)

Ne sont **PAS** considérés comme des exemple positifs (i.e lié à un événement de référence, annoté 0) les documents :

— Dont ni le titre ni le chapeau ne permet d'identifier clairement une mention de l'événement de référence, par exemple :

(A.8) a. Référence AFP :
Wildfires rampage across northern California.
Firefighters on Monday battled devastating wildfires that have reduced hundreds of homes to smoldering ruins and threatened California's renowned wine region.

b. Document valide pour une annotation de valeur 0 :
10 Things to Know for Tuesday. — Since starting Saturday, the blaze has consumed more than 95 square miles, destroyed hundreds of homes and forced thousands of residents to flee.

— Dont le contenu n'a rien à voir avec le document de référence, c'est-à-dire qu'il ne remplit aucune condition évoquée pour l'annotation 2 ou 1).

Exemples complémentaires :

(A.9) a. Référence AFP :
Wildfires rampage across northern California.
Firefighters on Monday battled devastating wildfires that have reduced

hundreds of homes to smoldering ruins and threatened California's renowned wine region.

b. Document valide pour une annotation de valeur 0 :
The Latest : Woman who died in wildfire was disabled retiree. MIDDLETOWN, Calif.

(A.10) a. Référence AFP :
Truck bomb kills at least 33 in Baghdad market.
A truck bomb ripped through a market in a Shiite-majority area of north Baghdad on Thursday, killing at least 33 people, security and medical officials said.

b. Document valide pour une annotation de valeur 0 :
ISIS Claims Responsibility For Baghdad Market Bombing.
Audio for this story from Morning Edition will be available at approximately 9 :00 a.m.

B

Volumétrie des sources d'articles Web utilisées

La période temporelle couverte s'étend de 2013 à 2016.

URL source	Nombre d'articles rattachés
news.yahoo.com	725 971
www.sfgate.com	546 135
www.washingtontimes.com	348 102
www.huffingtonpost.com	298 027
www.bbc.co.uk	195 358
abcnews.go.com	192 957
feeds.reuters.com	171 215
feeds.washingtonpost.com	170 988
telegraph.feedsportal.com	134 495
feeds.nydailynews.com	121 530
rss.nytimes.com	99 368
www.yahoo.com	98 626
rss.feedsportal.com	94 545
www.cbsnews.com	78 266
www.npr.org	67 628
www.nytimes.com	61 726
rss.cnn.com	55 636
feeds.abcnews.com	54 281
feeds.nbcnews.com	49 619
rssfeeds.usatoday.com	45 137
www.reuters.com	43 619
www.engadget.com	42 211
www.nydailynews.com	41 338
feeds.huffingtonpost.com	40 286
www.theverge.com	38 779
www.independent.co.uk	36 814
feedproxy.google.com	31 336
www.washingtonpost.com	30 931
feeds.cbsnews.com	26 547
online.wsj.com	26 211
www.wsj.com	23 905
www.telegraph.co.uk	23 648
www.cnet.com	14 114
www.newsweek.com	13 894
time.com	13 099
edition.cnn.com	12 716
www.nbcnews.com	12 520
www.zdnet.com	11 495
169 autres sources	20 205

C

Tableaux des accords inter-annotateurs pour l'annotation des instances d'événements

C.1 Kappas de Cohen

Identifiant de l'instance	Paire d'annotateurs	κ de Cohen
DRF81	(1, 2)	0,313
	(1, 3)	0,0
	(2, 3)	0,0
EIK25	(1, 2)	0,581
	(1, 3)	0,491
	(2, 3)	0,519
PJB88	(1, 2)	0,865
	(1, 3)	1,0
	(2, 3)	0,865
QML65	(1, 2)	0,677
	(1, 3)	0,639
	(2, 3)	0,930
EXC52	(1, 2)	1,0
	(1, 3)	1,0
	(2, 3)	1,0
BVV52	(1, 2)	1,0
	(1, 3)	0,929
	(2, 3)	0,929
GZY46	(1, 2)	0,876
	(1, 3)	1,0
	(2, 3)	0,876
HIM13	(1, 2)	0,609
	(1, 3)	0,082
	(2, 3)	0,074
FCM41	(1, 2)	1,0
	(1, 3)	1,0
	(2, 3)	1,0
TEE75	(1, 2)	0,743
	(1, 3)	0,868
	(2, 3)	0,868
QZW37	(1, 2)	0,419
	(1, 3)	0,224
	(2, 3)	0,066
TEU16	(1, 2)	0,0
	(1, 3)	0,0
	(2, 3)	0,407

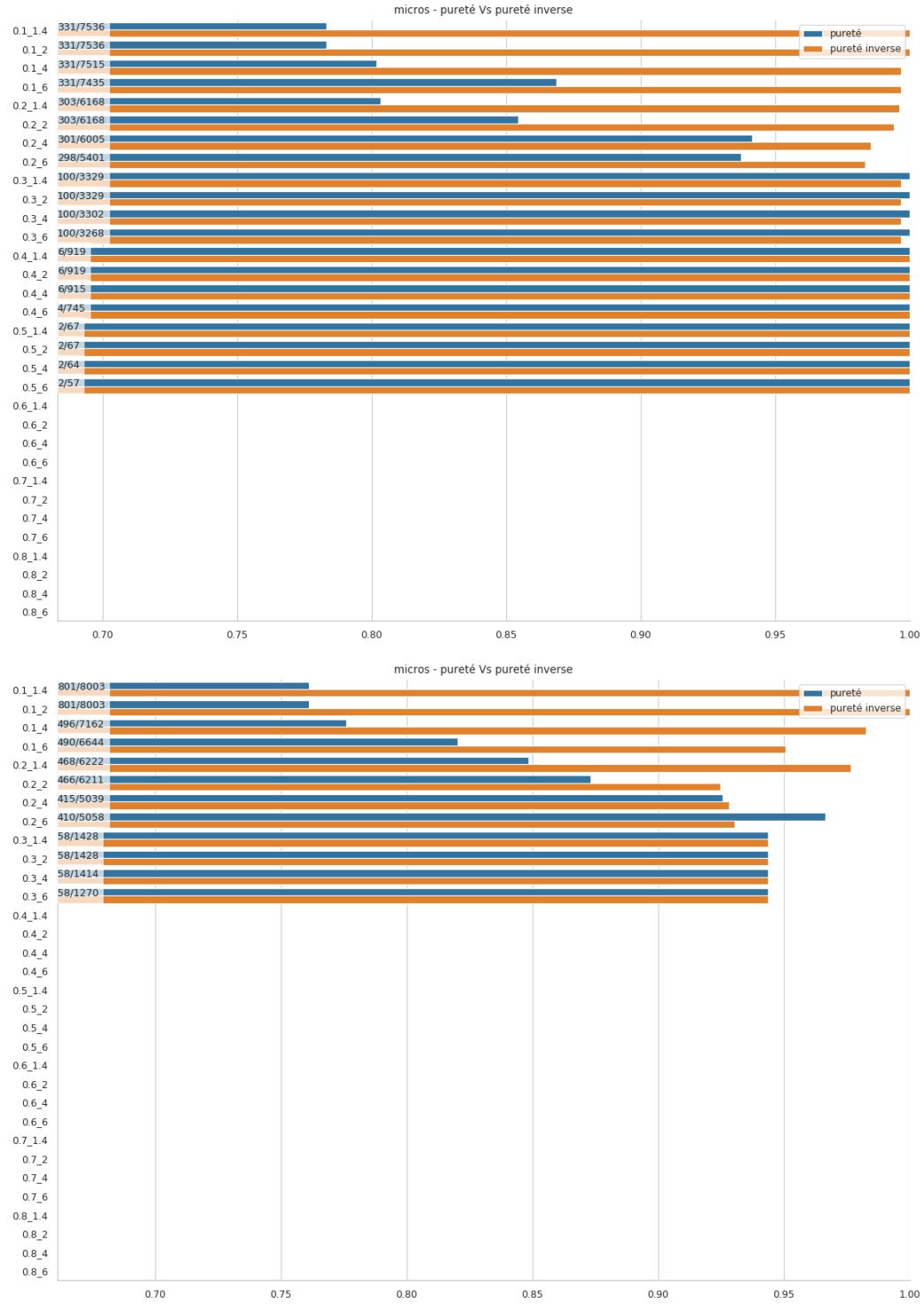
C.2 Kappas de Fleiss

Identifiant de l'instance	κ de Fleiss
BVV52	0,951
DRF81	0,102
FCM41	1,0
HIM13	0,208
PJB88	0,910
QML65	0,742
QZW37	0,110
TEE75	0,826
TEU16	0,116
EIK25	0,524
EXC52	1,0
GZY46	0,917

D

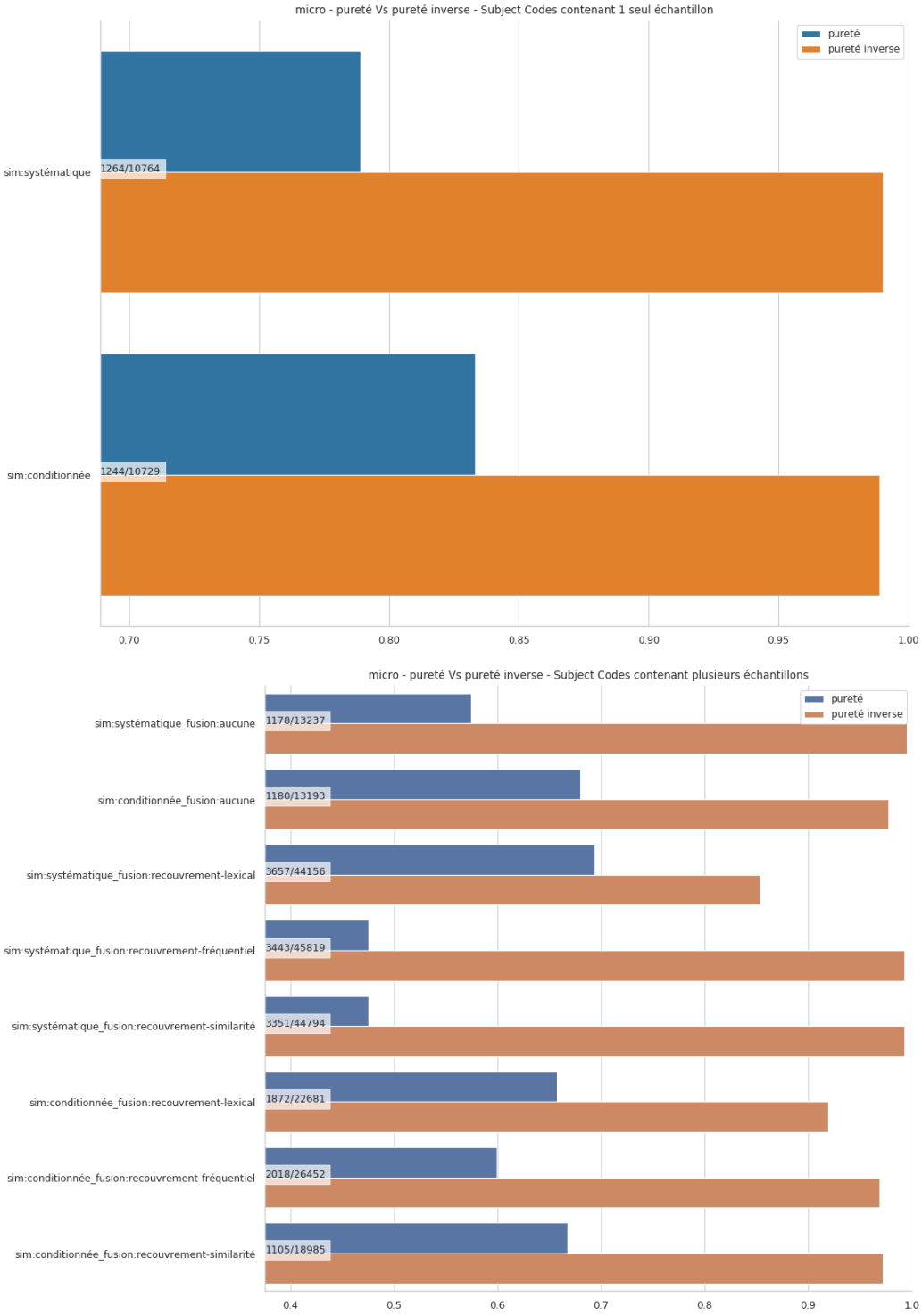
Résultats d'évaluation de la micro-pureté pour la phase d'induction de types

D.1 Résultats pour la baseline



Note : Comme pour la figure 4.5 en partie 4.4.4, l'histogramme du haut couvre les résultats sur les Subject Codes contenant moins de 100 instances et celui du bas les Subject Codes contenant plus de 100 instances.

D.2 Résultats pour notre système



Note : Comme pour la figure 4.6 en partie 4.4.4, l'histogramme du haut couvre les résultats sur les Subject Codes contenant moins de 100 instances et celui du bas les Subject Codes contenant plus de 100 instances.

Bibliographie

AGUILAR, Jacqueline, Charley BELLER, Paul MCNAMEE et Benjamin VAN DURME (2014). "A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP, and FrameNet Annotation Standards". In : *Proceedings of the 2nd Workshop on EVENTS : Definition, Detection, Coreference, and Representation, pages 45–53, Baltimore, Maryland, USA, June 22-27, 2014. c 2014 Association for Computational Linguistics* (cf. p. 17).

AHN, Natalie (2017). "Inducing Event Types and Roles in Reverse : Using Function to Discover Theme". In : *Proceedings of the Events and Stories in the News Workshop*, p. 66–76 (cf. p. 33, 35, 65).

ALLAN, James (2002). *Topic Detection and Tracking : Event-based Information Organization*. Kluwer Academic (cf. p. 17, 18).

AZZOPARDI, Joel et Christopher STAFF (2012). "Incremental Clustering of News Reports". In : *Algorithms* (cf. p. 24, 40).

BAKER, Collin F., Charles J. FILLMORE et John B. LOWE (1998). "The Berkeley FrameNet Project". In : *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. ACL '98/COLING '98. Montreal, Quebec, Canada : Association for Computational Linguistics, p. 86–90 (cf. p. 88).

BANKO, Michele, Michael J. CAFARELLA, Stephen SODERLAND, Matt BROADHEAD et Oren ETZIONI (2007). "Open Information Extraction from the Web". In : *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07), pp. 2670–2676, Hyderabad, India, 2007.* (Cf. p. 21).

BAYARDO, Roberto J., Yiming MA et Ramakrishnan SRIKANT (2007). "Scaling up all pairs similarity search". In : *16^{th} International World Wide Web Conference (WWW'07)*. Banff, Alberta, Canada, p. 131–140 (cf. p. 40).

BLEI, David M., Andrew Y. NG et Michael I. JORDAN (2003). "Latent Dirichlet Allocation". In : *Journal of Machine Learning Research 3 (2003) 993-1022* (cf. p. 25).

BORGWARDT, Karsten Michael (2007). "Graph kernels". Thèse de doct. lmu (cf. p. 24).

CHAMBERS, Nathanael (2013). "Event Schema Induction with a Probabilistic Entity-Driven Model". In : *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1797–1807, Seattle, Washington, USA, 18-21 October 2013. c 2013 Association for Computational Linguistics* (cf. p. 29–31).

CHAMBERS, Nathanael et Dan JURAFSKY (2008). "Unsupervised Learning of Narrative Event Chains". In : *Proceedings of ACL 2008* (cf. p. 22).

CHAMBERS, Nathanael et Dan JURAFSKY (2009). “Unsupervised Learning of Narrative Schemas and their Participants”. In : *Proceedings of ACL 2009* (cf. p. 23).

– (2011). “Template-Based Information Extraction without the Templates”. In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Portland, Oregon, USA, p. 976–986 (cf. p. 28, 29, 33, 35, 42, 111).

CHEN, Dawn, Joshua C. PETERSON et Thomas L. GRIFFITHS (2017). *Evaluating vector-space models of analogy*. arXiv : 1705.04416 [cs.CL] (cf. p. 109).

CHEN, Yanping, Qinghua ZHENG, Feng TIAN et al. (2017). “Exploring Open Information via Event Network”. In : *Natural Language Engineering* 24, p. 1–22 (cf. p. 25).

CHEUNG, Jackie Chi Kit, Hoifung POON et Lucy VANDERWENDEN (2013). “Probabilistic Frame Induction”. In : *Proceedings of NAACL-HLT 2013*. Atlanta, Georgia, USA, p. 837–846 (cf. p. 28, 35).

CHRISTENSEN, Janara, Stephen SODERLAND, Oren ETZIONI et al. (2010). “Semantic role labeling for open information extraction”. In : *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. Association for Computational Linguistics, p. 52–60 (cf. p. 22).

CONRAD, Jack G. et Michael BENDER (2016). “Semi-Supervised Events Clustering in News Retrieval”. In : *Proceedings of the NewsIR’16 Workshop at ECIR, Padua, Italy* (cf. p. 24, 37, 40).

DEL CORRO, Luciano et Rainer GEMULLA (2013). “Clausie : clause-based open information extraction”. In : *Proceedings of the 22nd international conference on World Wide Web*. ACM, p. 355–366 (cf. p. 22).

DENG, Lingjia et Janyce WIEBE (2015). “Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models”. In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal : Association for Computational Linguistics, p. 179–189 (cf. p. 19).

DEVLIN, Jacob, Ming-Wei CHANG, Kenton LEE et Kristina TOUTANOVA (2018). “Bert : Pre-training of deep bidirectional transformers for language understanding”. In : *arXiv preprint arXiv :1810.04805* (cf. p. 109).

DING, Haibo et Ellen RILOFF (2018). “Human Needs Categorization of Affective Events Using Labeled and Unlabeled Data”. In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana : Association for Computational Linguistics, p. 1919–1929 (cf. p. 19).

DODDINGTON, George R., Alexis MITCHELL, Mark A. PRZYBOCKI et al. (2004). “The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation”. In : *LREC* (cf. p. 16, 17).

DONGEN, Stijn van (2000). “Graph Clustering by Flow Simulation”. Thèse de doct. University of Utrecht (cf. p. 41).

FERRET, Olivier (1998). “ANTHAPSI : un système d’analyse thématique et d’apprentissage de connaissances pragmatiques fondé sur l’amorçage”. Thèse de doct. (cf. p. 18, 65).

FILATOVA, Elena, Vasileios HATZIVASSILOGLOU et Kathleen MCKEOWN (2006). “Automatic Creation of Domain Templates”. In : *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia, p. 207–214 (cf. p. 6, 27, 65).

FINKEL, Jenny Rose, Trond GRENAGER et Christopher MANNING (2005). “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”. In : *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL ’05. Ann Arbor, Michigan : Association for Computational Linguistics, p. 363–370 (cf. p. 44).

GANITKEVITCH, Juri, Benjamin VAN DURME et Chris CALLISON-BURCH (2013). “PPDB : The Paraphrase Database”. In : *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT ’13*. Atlanta, Georgia, USA, p. 758–764 (cf. p. 45).

GAO, Yuyang et Liang ZHAO (2018). “Incomplete Label Multi-Task Ordinal Regression for Spatial Event Scale Forecasting”. In : p. 2999–3006 (cf. p. 19).

GLAVAŠ, Goran et Jan ŠNAJDER (2013). “Recognizing Identical Events with Graph Kernels”. In : *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria, p. 797–803 (cf. p. 24, 40, 60).

GRISHMAN, Ralph et Beth SUNDHEIM (1996). “Message Understanding Conference - 6 : A Brief History”. In : *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96), Copenhagen, August 1996*, p. 466–471 (cf. p. 16).

GUI, Lin, Dongyin WU, Ruifeng XU, Qin LU et Yu ZHOU (2016). “Event-Driven Emotion Cause Extraction with Corpus Construction”. In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas : Association for Computational Linguistics, p. 1639–1649 (cf. p. 19).

HASHIMOTO, Chikara, Kentaro TORISAWA, Julien KLOETZER et Jong-Hoon OH (2015). “Generating Event Causality Hypotheses Through Semantic Relations”. In : *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas : AAAI Press, p. 2396–2403 (cf. p. 20).

HUANG, Lifu, Taylor CASSIDY, Feng XIAOCHENG et al. (2016). “Liberal Event Extraction and Event Schema Induction”. In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin, Germany, p. 258–268 (cf. p. 32).

KRUENGKRAI, Canasai, Kentaro TORISAWA, Chikara HASHIMOTO et al. (2017). “Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks”. In : *Thirty-First AAAI Conference on Artificial Intelligence* (cf. p. 20).

KUHN, Harold W (1955). “The Hungarian method for the assignment problem”. In : *Naval research logistics quarterly* 2.1-2, p. 83–97 (cf. p. 70).

KUZEY, Erdal, Jilles VREEKEN et Gerhard WEIKUM (2014). “A Fresh Look on Knowledge Bases : Distilling Named Events from News”. In : *Proceedings of CIKM 2014, Shanghai, China* (cf. p. 30, 81).

LEBOWITZ, Michael (1983). “Generalization from Natural Language Text”. In : *Cognitive Science* 7.1, p. 1–40 (cf. p. 15).

LEVY, Omer, Yoav GOLDBERG et Ido DAGAN (2015). “Improving Distributional Similarity with Lessons Learned from Word Embeddings”. In : *Transactions of the Association for Computational Linguistics* 3, p. 211–225 (cf. p. 109).

LI, Wei, Lei HE et Hai ZHUGE (2016). “Abstractive news summarization based on event semantic link network”. English. In : *The 26th International Conference on Computational Linguistics*. Association for Computational Linguistics (cf. p. 19).

LINGUISTIC DATA CONSORTIUM (2005). *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. `https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf`. [En ligne au 28 Mars 2019] (cf. p. 17).

LIU, Xiao, Heyan HUANG et Yue ZHANG (2019). “Open Domain Event Extraction Using Neural Latent Variable Models”. In : *Proceedings of Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (cf. p. 31, 35).

MARNEFFE, Marie-Catherine de, Bill MACCARTNEY et Christopher D. MANNING (2006). “Generating Typed Dependency Parses from Phrase Structure Parses”. In : *LREC* (cf. p. 44).

MIKOLOV, Tomas, Ilya SUTSKEVER, Kai CHEN, Greg S CORRADO et Jeff DEAN (2013). “Distributed representations of words and phrases and their compositionality”. In : *Advances in neural information processing systems*, p. 3111–3119 (cf. p. 109).

MINSKY, Marvin (1974). “A Framework for Representing Knowledge”. In : *The Psychology of Computer Vision* (cf. p. 15).

MODI, Ashutosh (2016). “Event Embeddings for Semantic Script Modeling”. In : *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany : Association for Computational Linguistics, p. 75–83 (cf. p. 111).

MOONEY, Raymond (1985). “Learning Schemata for Natural Language Processing”. In : Actes Ninth International Joint Conference on Artificial Intelligence, Los Angeles, p. 681–687 (cf. p. 15).

NANNI, Federico, Simone Paolo PONZETTO et Laura DIETZ (2017). “Building Entity-centric Event Collections”. In : *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. JCDL ’17. Toronto, Ontario, Canada : IEEE Press, p. 199–208 (cf. p. 26, 37).

NGUYEN, Kiem-Hieu, Xavier TANNIER, Olivier FERRET et Romaric BESANÇON (2015). “Generative Event Schema Induction with Entity Disambiguation”. In : *Proceedings of ACL 2015* (cf. p. 30, 31, 33, 35).

NGUYEN, Kiem-Hieu, Xavier TANNIER et Véronique MORICEAU (2014). “Ranking Multidocument Event Descriptions for Building Thematic Timelines”. In : *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*. Dublin, Ireland : Dublin City University et Association for Computational Linguistics, p. 1208–1217 (cf. p. 19).

NIMISHAKAVI, Madhav, Manish GUPTA et Partha TALUKDAR (2018). “Higher-order Relation Schema Induction using Tensor Factorization with Back-off and Aggregation”. In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Melbourne, Australia : Association for Computational Linguistics, p. 1575–1584 (cf. p. 81).

ORR, John Walker, Prasad TADEPALLI, Janardhan Rao DOPPA, Xiaoli FERN et Thomas G DIETTERICH (2014). “Learning Scripts as Hidden Markov Models”. In : *Twenty-Eighth AAAI Conference on Artificial Intelligence* (cf. p. 15).

OUYANG, Jessica et Kathy MCKEOWN (2019). “Neural Network Alignment for Sentential Paraphrases”. In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy : Association for Computational Linguistics, p. 4724–4735 (cf. p. 109).

PADRÓ, Lluís, Zeljko AGIC, Xavier CARRERAS et al. (2014). “Language Processing Infrastructure in the XLike Project”. In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 26-31 May 2014*. Reykjavik, Iceland (cf. p. 23).

PENNINGTON, Jeffrey, Richard SOCHER et Christopher MANNING (2014). “Glove : Global vectors for word representation”. In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543 (cf. p. 109).

PETERS, Matthew, Mark NEUMANN, Mohit IYYER et al. (2018). “Deep Contextualized Word Representations”. In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana : Association for Computational Linguistics, p. 2227–2237 (cf. p. 109).

PICHOTTA, Karl et Raymond J MOONEY (2016). “Learning Statistical Scripts with LSTM Recurrent Neural Networks”. In : *Thirtieth AAAI Conference on Artificial Intelligence* (cf. p. 15).

PUSTEJOVSKY, James, Kiyong LEE, Harry BUNT et Laurent ROMARY (2010). “ISO-TimeML : An International Standard for Semantic Annotation”. In : *LREC 2010*, p. 394–397 (cf. p. 19).

RADINSKY, Kira, Sagie DAVIDOVICH et Shaul MARKOVITCH (2012). “Learning causality for news events prediction”. In : *Proceedings of the 21st international conference on World Wide Web*. ACM, p. 909–918 (cf. p. 87).

REIMERS, Nils et Iryna GUREVYCH (2019). “Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks”. In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China : Association for Computational Linguistics, p. 3980–3990 (cf. p. 109).

ROSPOCHER, Marco, Marieke van ERP, Piek VOSSEN et al. (2016). “Building Event-centric Knowledge Graphs from News”. In : *Web Semant.* 37.C, p. 132–151 (cf. p. 26, 40).

ROY, Arpita, Youngja PARK, Taesung LEE et Shimei PAN (2019). “Supervising Unsupervised Open Information Extraction Models”. In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 728–737 (cf. p. 22).

RUDNIK, Charlotte, Thibault EHRHART, Olivier FERRET et al. (2019). “Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata”. In : *Proceedings of the Wiki Workshop 2019 (The Web Conference)*. San Francisco, USA (cf. p. 11, 49, 73, 94, 110).

RUPPENHOFER, Josef, Michael ELLSWORTH, Myriam SCHWARZER-PETRUCK, Christopher R JOHNSON et Jan SCHEFFCZYK (2006). “FrameNet II : Extended theory and practice”. In : (cf. p. 88).

SCHANK, Roger C. et Robert P. ABELSON (1977). *Scripts, plans, goals and understanding*. Lawrence Erlbaum (cf. p. 15, 22).

SEKINE, Satoshi (2006). “On-Demand Information Extraction”. In : *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia : Association for Computational Linguistics, p. 731–738 (cf. p. 21).

SHA, Lei, Sujian LI, Baobao CHANG et Zhifang SUI (2016). “Joint Learning Templates and Slots for Event Schema Induction”. In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics (cf. p. 31, 35).

SHINYAMA, Yusuke et Satoshi SEKINE (2006). “Preemptive Information Extraction using Unrestricted Relation Discovery”. In : *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. New York City, USA : Association for Computational Linguistics, p. 304–311 (cf. p. 21).

SONG, Zhiyi, Ann BIES, Stephanie STRASSEL et al. (2015). “From Light to Rich ERE : Annotation of Entities, Relations, and Events”. In : *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, pages ,Denver, Colorado, June 4, 2015*. P. 89–98 (cf. p. 17).

SPRUGNOLI, Rachele et Sara TONELLI (2017). “One, no one and one hundred thousand events : Defining and processing events in an inter-disciplinary perspective”. In : *Natural Language Engineering* 23.4, p. 485–506 (cf. p. 23, 35).

STRÖTGEN, Jannik et Michael GERTZ (2012). “Multilingual and cross-domain temporal tagging”. In : *Language Resources and Evaluation* 47 (cf. p. 19).

SULTAN, Md Arafat, Steven BETHARD et Tamara SUMMER (2014). “Back to Basics for Monolingual Alignment : Exploiting Word Similarity and Contextual Evidence”. In : *Transactions of the Association for Computational Linguistics (TACL)* 2, p. 219–230 (cf. p. 44, 69, 75).

SUN, Rui, Yue ZHANG, Meishan ZHANG et Donghong JI (2015). “Event-Driven Headline Generation”. In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. Beijing, China : Association for Computational Linguistics, p. 462–472 (cf. p. 19).

SUNDHEIM, Beth M et Nancy A CHINCHOR (1993). “Survey of the message understanding conferences”. In : *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, p. 56–60 (cf. p. 16).

TANNIER, Xavier (2014). *Traitement des événements et ciblage d’information*. Habilitation à Diriger des Recherches (HDR) (cf. p. 28, 35).

TANNIER, Xavier, Véronique MORICEAU, Béatrice ARNULPHY et Ruixin HE (2012). “Evolution of Event Designation in Media : Preliminary Study”. In : *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey (cf. p. 28).

TOUILEB, Samia et Katherine DUARTE (2016). “Getting to know large newsflows : Automatically induced information structures as keyphrases for news content analysis”. In : p. 35–40 (cf. p. 19).

TOUTANOVA, Kristina, Dan KLEIN, Christopher D. MANNING et Yoram SINGER (2003). “Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network”. In : *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL ’03. Edmonton, Canada : Association for Computational Linguistics, p. 173–180 (cf. p. 44).

VASWANI, Ashish, Noam SHAZEER, Niki PARMAR et al. (2017). “Attention is All you Need”. In : *Advances in Neural Information Processing Systems 30*. Sous la dir. d’I. GUYON, U. V. LUXBURG, S. BENGIO et al. Curran Associates, Inc., p. 5998–6008 (cf. p. 109).

YUAN, Quan, Xiang REN, Wenqi HE et al. (2018). “Open-Schema Event Profiling for Massive News Corpora”. In : *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM ’18. Torino, Italy : ACM, p. 587–596 (cf. p. 33).

Titre : Induction non-supervisée de schémas d'événements à partir de textes journalistiques

Mots clés : Recherche d'Information, Data-journalisme, Clustering

Résumé : L'événement est un concept central dans plusieurs tâches du Traitement Automatique des Langues, en dépit de l'absence d'une définition unifiée de ce que recouvre cette notion. Le traitement des événements s'est structuré sous l'égide des campagnes d'évaluation MUC (*Message Understanding Conference*), qui fournissaient des structures de référence appelées schémas (*templates* en anglais), se présentant sous la forme d'un titre et d'une collection d'arguments (*slots*), chacun représentant un élément caractéristique de l'événement décrit (par exemple l'*épicentre* d'un séisme). La création de ces schémas requiert une connaissance experte et est donc longue, coûteuse et difficile à étendre à un large ensemble de domaines de spécialité.
En parallèle de ces travaux, la quantité de données produites par les individus et les organisations a crû de manière exponentielle, ouvrant des perspectives applicatives inédites. Cette croissance a notamment favorisé l'essor d'un nouveau paradigme journalistique appelé journalisme de données (*data-journalism* en anglais).
Le présent travail se propose d'induire, à partir d'un grand volume de texte journalistique et sans supervision, des représentations synthétiques d'événements journalistiques comparables aux *templates* des campagnes MUC, dans l'objectif de faciliter l'exploitation de grandes masses de données par des journalistes des données. Pour ce faire, nous suivons une approche ascendante divisée en trois grandes étapes. Dans la première étape, nous groupons ensemble les nombreuses mentions textuelles relatant la même réalisation d'un événement, identifiée dans le temps et l'espace et appelée instance. La deuxième étape vise à s'abstraire des caractéristiques spatio-temporelles de chaque instance pour les grouper en grands types d'événements. Enfin, la dernière étape de cette contribution vise à extraire les éléments caractéristiques de chaque type d'événement induit afin d'en proposer une représentation synthétique assimilable à un schéma d'événement.

Title : Unsupervised event schemas induction from journalistic texts

Keywords : Information Retrieval, Data-journalism, Clustering

Abstract : Events are central in many Natural Language Processing tasks, despite the lack of a unified definition for the concept. The field of event processing took off with the MUC evaluation campaigns that provided participants with reference structures called templates. These templates were composed of a title (the name of the event) and several slots, i.e specific and atomic pieces of data about the event. Creating these templates is an expert task and therefore costly, painstaking and hard to extend to new domains.
Meanwhile, the amount of data produced by individuals and organizations has grown exponentially, opening unprecedented perspectives of applications. In the journalistic domain, it fueled the development of a new paradigm called data-journalism.
In this work, we aim at inducing synthetic representations of events from large textual journalistic corpora. These representations would be comparable to MUC templates and used by data-journalists to explore large textual news datasets. To this end, we propose a bottom-up approach composed of three main steps. The first step clusters several textual mentions of a same particular event (i.e tied to a time and place) to identify distinct *instances*. The second step groups these instances together based on more abstract features to infer event types. Finally, the third and last step extracts the most salient elements of each type to produce the synthetic, template-like structure we are looking for.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France