Traitement des événements et ciblage d'information

Xavier Tannier

Traitement des événements et ciblage d'information

MÉMOIRE D'HABILITATION À DIRIGER LES RECHERCHES

Spécialité INFORMATIQUE

présenté par

Xavier TANNIER

le 18 juin 2014

au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS, Orsay) devant un jury composé de

Béatrice Daille	Professeure à l'Université de Nantes rapporteure
Marie-Francine Moens	Professeure à l'Université Catholique de Louvain rapporteure
Pascale Sébillot	Professeure à l'INSA de Rennes rapporteure
Catherine Berrut	Professeure à l'Université Joseph Fourier de Grenoble examinatrice
Patrice Bellot	Professeur à Aix-Marseille Université examinateur
Sophie Rosset	Directrice de recherches au CNRS, LIMSI examinatrice
Pierre Zweigenbaum	Directeur de recherches au CNRS, LIMSI parrain

Table des matières

Remerciements

Mes remerciements vont en premier lieu à Béatrice Daille, Marie-Francine Moens et Pascale Sébillot qui ont accepté la tâche de rapporter sur ce mémoire, ainsi que Catherine Berrut, Sophie Rosset et Patrice Bellot pour leur présence dans ce jury. Merci également à Pierre Zweigenbaum pour m'avoir accompagné et conseillé tout au long de ce travail.

Il est banal de le dire, mais il serait bien ingrat de le taire : le parcours d'un chercheur est une affaire de rencontre. En ce qui me concerne, il s'est souvent agi de rencontres avec des personnes qui m'ont fait confiance alors que je n'avais encore rien fait pour le mériter, au premier rang desquelles figurent Philippe Muller, Shlomo Geva, Frédérique Segond. Une bonne dose de chance n'est jamais de trop pour se frayer un chemin dans le monde de la recherche, la mienne a été de vous croiser.

Une autre chance a été d'être accueilli au LIMSI, dans l'équipe ILES, et l'ambiance qui y règne à tous les niveaux fait qu'on y apprend et grandit plus vite. Je me garderai bien d'une énumération de noms forcément incomplète, mais soyez-en tous remerciés.

Enfin, l'enseignement à l'Université Paris-Sud tient une place importante dans mon quotidien comme dans mon intérêt pour le métier que je fais. L'administration de l'enseignement, quant à elle... doit être assurée. Avoir des collègues efficaces, dévoués à la cause et sympathiques permet d'affronter les difficultés avec sérénité et bonne humeur, dans ces tâches parfois passionnantes et parfois ingrates.

Avant-propos

Ce mémoire présente une synthèse de mes activités dans le domaine de l'extraction et de la recherche d'information, accompagnée d'une réflexion sur les pistes de recherche qui peuvent en découler.

Après un mémoire de DEA à l'IRIT sur l'extraction des relations temporelles dans les textes, que je mentionne car il s'est avéré d'une influence pour la suite que j'aurais eu bien du mal à mesurer à l'époque, ces activités ont débuté par des travaux orientés vers la question de la recherche d'information ciblée. Le ciblage a d'abord consisté en l'utilisation de la structure des documents XML, durant ma thèse à l'École des Mines de Saint-Etienne, puis, après un bref retour vers les relations temporelles pendant mon post-doc au centre de recherche de Xerox à Grenoble, en un travail sur les systèmes dits de questions-réponses, dans l'équipe ILES du LIMSI.

Peu à peu, je me suis ensuite focalisé sur les aspects temporels et événementiels des textes, que ce soit pour du ciblage d'information ou pour de l'agrégation, de la contextualisation, de la synthèse d'information. J'ai également conservé quelques activités plus directement liées à la recherche d'information.

Sans aborder tous les méandres de ce parcours, j'en ai extrait ce que j'estime en être la substantifique moëlle (ou, en tout cas, la moëlle) pour parvenir à ce mémoire qui, je l'espère, permettra autant au lecteur de se projeter vers l'avenir que de découvrir le passé. Une annexe présente de façon factuelle l'ensemble de mes activités d'enseignement, d'encadrement et de recherche.

Guide de lecture

Organisé en deux parties consacrées au traitement des événements d'une part, et au ciblage d'information d'autre part, ce document comprend cinq chapitres principaux et une conclusion.

Le premier chapitre, entièrement original, introduit la notion d'événement à travers les diverses spécialités du traitement automatique des langues qui s'en sont préoccupées. Il propose un survol global des différents modes de représentation des événements choisis par des chercheurs d'horizons divers. Il instaure également un fil rouge pour l'ensemble de la première partie, à travers la présentation de différents axes de recherche sur lesquels nous nous appuierons pour contextualiser les chapitres suivants.

Les quatre chapitres restants reprennent des travaux effectués depuis mon doctorat. Les deux premiers distinguent deux grandes classes de travaux autour des événements, deux grandes visions que nous avons nommées, pour la première, "l'événement dans le texte", et pour la seconde, "l'événement dans le monde". Dans la première (chapitre 2), nous considérons l'événement comme la désignation linguistique de quelque

chose qui se passe, et nous tentons d'une part d'identifier ces désignations dans les textes (section 2.1), et d'autre part d'induire les relations temporelles existants entre ces événements, que ce soit dans des textes journalistiques (section 2.2) ou médicaux (section 2.3). Nous réfléchissons enfin à une métrique d'évaluation adaptée à ce type d'informations (section 2.4).

Pour ce qui est de l'"événement dans le monde" (chapitre 3), nous envisageons plus l'événement tel qu'il est perçu par le citoyen, et nous proposons plusieurs approches originales pour aider celui-ci à mieux appréhender la quantité écrasante d'événements dont il prend connaissance chaque jour : les chronologies thématiques (section 3.1), les fils temporels (section 3.2), et une approche automatisée du journalisme de données (section 3.3).

Chacune de ces visions se nourrit de l'autre et nous verrons d'ailleurs dans les parties consacrées aux perspectives qu'un des buts que nous nous fixons est d'abolir partiellement la frontière qui les séparent.

La deuxième partie peut se lire de façon indépendante de la première. Elle contient deux chapitres revenant sur des travaux en lien avec le ciblage d'information. Dans le chapitre 4, nous décrivons nos travaux sur les systèmes de questions-réponses, dans lesquels nous avons eu recours à l'analyse syntaxique pour aider à justifier les réponses trouvées à une question en langage naturel. Enfin, le chapitre 5 aborde le sujet de la collecte thématique de documents sur le Web, dans le but de créer automatiquement des corpus et des lexiques spécialisés.

Notons que si nous nous sommes beaucoup préoccupé de l'évaluation des approches mises en œuvre, et ce dans l'ensemble des travaux décrits, ce mémoire ne présente les résultats obtenus en détails que lorsque la méthodologie suivie pour l'évaluation est elle-même originale et fait partie des innovations que nous souhaitons mettre en avant. Dans le cas contraire, nous orientons le lecteur vers les articles publiés sur chaque thème, qui contiennent tous les détails liés aux résultats quantitatifs. De la même façon, nous nous attardons très peu sur les caractéristiques des corpus manipulés (taille, répartition pour l'apprentissage, etc.).

Enfin, le chapitre 6 apporte une conclusion et revient sur les perspectives associées aux travaux présentés.

Première partie

L'événement dans le texte et dans le monde

Traitements des événements et des informations temporelles

Chapitre 1

L'événement : définitions, représentations, intérêts

Tout scientifique aime définir correctement les objets qu'il manipule. Dans des domaines comme la philosophie ou les sciences du language, la question de la définition de la notion d'événement a fait l'objet de réflexions poussées. De ces réflexions émergent les idées de "*changement d'état du monde*" [96], de localisation spatio-temporelle [154], d'importance pour une communauté donnée [117], ainsi que le lien avec les notions de contexte et de référence [36], qui prennent toute leur importance en extraction d'information.

En traitement automatique des langues (TAL), en revanche, où de nombreux travaux sont pourtant consacrés aux événements, cet effort de définition est très rarement fait. On se penche certes sur l'environnement de l'événement, on crée des taxinomies et des typologies, le plus souvent dans un but applicatif précis. Mais la définition elle-même de l'événement ne dépasse finalement pas le lieu commun :

> « On peut dire, grossièrement, que les événements sont des choses qui arrivent à un certain moment. »
> [43, in Essai n°11] [1]

Le TAL est une science appliquée, et nous pensons que si la question de la définition de l'événement est globalement mise de côté, c'est parce qu'une autre question, plus directement liée aux applications et ayant des conséquences plus immédiates et plus visibles, prévaut : celle de la représentation.

En effet, un bref parcours de la littérature sur le traitement des événements en TAL suffit à s'en persuader : la façon de désigner et de représenter l'événement varie du tout au tout selon les sous-domaines, les applications, les types de corpus exploités. Les communautés concernées n'ont que peu de contact les unes avec les autres, et leur unique point commun, en plus de celui de traiter du langage, ne semble finalement être que de manipuler "des choses qui arrivent à un certain moment".

Ce chapitre ne se veut pas une introduction de l'état de l'art en matière de traitement des événements dans la langue écrite : les travaux liés plus directement avec les thématiques de recherche abordées dans ce mémoire seront présentés au fil des chapitres suivants. Il propose plutôt un survol global des différents modes de représentation des événements choisis par des chercheurs d'horizons divers. Il est bien entendu que les

1. Traduit en français en 1993 [44].

représentations choisies sont indissociables du type de besoin d'information que l'on souhaite satisfaire (questions précises, résumé, acquisition de connaissances...) et par conséquent de l'application visée. On choisit généralement la représentation en fonction de l'application, et pas l'inverse. Pourtant, aborder les choses sous l'angle de la représentation permet d'élargir l'horizon de chaque domaine. Un article scientifique spécialisé, devant présenter une approche en quelques pages, se focalise de façon naturelle sur les objectifs visés ; un chercheur en quête d'état de l'art et de nouvelles perspectives risque de passer son chemin si le contexte décrit lui semble trop éloigné de ses centres d'intérêt. Ce chapitre est donc l'occasion de changer de point de vue et d'explorer toutes les sauces auxquelles l'événement peut être mangé. Les aspects applicatifs sont bien entendu mentionnés en toile de fond.

Nous abordons donc les représentations diverses de la notion d'événement, à travers quatre grandes classes allant de la représentation la plus succinte, la plus précise et la plus attachée au texte, jusqu'à la plus relâchée, la moins contrainte et la plus abstraite. L'événement prendra donc, tour à tour, la forme de la simple annotation d'un mot dans une phrase (section 1.1), d'un schéma informationnel relationnel (section 1.2), d'un sac de mots (section 1.3) et enfin d'un pic statistique (section 1.4).

En parcourant ces différentes visions de l'événement, nous pourrons noter qu'il existe une certaine dualité dans la notion de "représentation", qui peut parfois prêter à confusion. On oscille parfois entre une représentation axée sur l'*information textuelle* (les mots que l'auteur d'un texte choisit pour décrire l'événement) et une représentation axée sur les *connaissances*, dans laquelle on manipule plutôt des concepts reliés entre eux par des relations. Une simple annotation de texte appartient plutôt à la première catégorie. Par exemple, selon la norme TimeML (section 1.1), un événement est désigné par un mot dans une phrase. En revanche, si l'on s'abstrait du texte pour remplir des schémas ou faire des déductions statistiques à partir d'une certaine quantité de mentions d'un événement, on se dirige vers la représentation des connaissances, qui ne s'appuie pas sur du texte pour montrer les choses. Si la frontière est floue entre ces deux modes, c'est parce que dans notre domaine, la représentation des connaissances est finalement bel et bien assurée par du texte (il faut bien désigner un concept avec un mot ou un ensemble de mots), et que ce texte est souvent issu d'un document.

Si cette distinction n'a pas une incidence capitale sur la compréhension des choses, il n'est toutefois pas inutile de la garder à l'esprit lorsque l'on s'attarde sur les modes de représentation.

1.1 Un mot dans la phrase et dans le document

Un domaine particulièrement dynamique ces dernières années est celui de la détection des événements et des expressions temporelles dans un texte [136, 103, 142, 35], puis l'établissement des relations temporelles (*avant*, *après*, *pendant*...) entre ces éléments [107, 127, 151, 50].

Ici, un événement est vu comme le mot qui en porte la désignation dans la phrase. Le plus souvent, il s'agit d'un verbe ou de la tête d'un groupe nominal, mais selon la norme TimeML [128], devenue le standard en la matière, il est tout à fait possible de considérer des adjectifs, ou même des valeurs numériques, comme des événements.

L'extraction d'information temporelle et événementielle s'articule alors traditionnellement autour des trois tâches suivantes, conduisant à une annotation de textes illustrée aux figures 1.1 et 1.2 :

1. Identification des mots désignant des événements dans les textes (balises `EVENT` dans le standard TimeML). Idéalement, identification des caractéristiques gram-

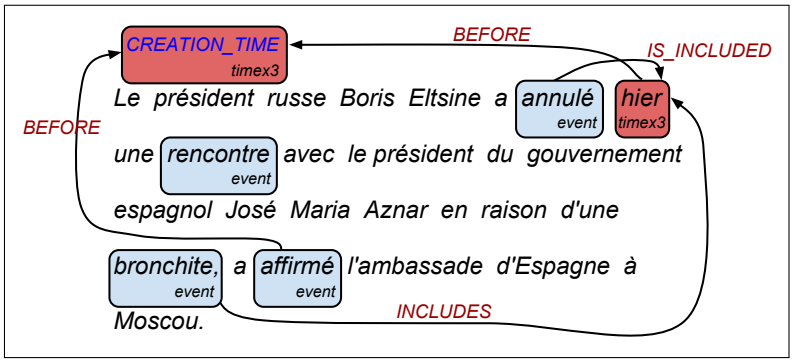

FIGURE 1.1: Exemple d'annotation des événements et des expressions temporelles, représentation graphique. La version textuelle annotée en XML est présentée à la figure 1.2.

```
<TEXT>
    <TIMEX3 functionInDocument="CREATION_TIME" tid="t1"
    type="DATE" value="1999-05-19"/>

    Le président russe Boris Eltsine a <EVENT class="I_ACTION" eiid="ei1"
    pos="VERB" pred="ANNULER" tense="PAST" vform="PASTPART">annulé</EVENT>
    <TIMEX3 temporalFunction="true" tid="t2" type="DATE" value="1999-05-18"
    valueFromFunction="tf17">hier</TIMEX3> une <EVENT class="OCCURRENCE"
    eiid="ei2" pos="NOUN" pred="RENCONTRE">rencontre</EVENT>
    avec le président du gouvernement espagnol Jose
    Maria Aznar en raison d'une <EVENT class="STATE"
    eiid="ei14" pos="NOUN" pred="BRONCHITE">bronchite</EVENT>, a
    <EVENT class="REPORTING" eiid="ei4" pos="VERB" pred="AFFIRMER" tense="PAST"
    vform="PASTPART">affirmé</EVENT> l'ambassade d'Espagne à Moscou.
</TEXT>
<TLINK lid="l1" relType="BEFORE" relatedToTime="t1" timeID="t2"/>
<TLINK eventInstanceID="ei1" lid="l4" relType="IS_INCLUDED" relatedToTime="t2"/>
<TLINK eventInstanceID="ei4" lid="l6" relType="BEFORE" relatedToTime="t1"/>
<TLINK eventInstanceID="ei14" lid="l10" relType="INCLUDES" relatedToTime="t2"/>
```

FIGURE 1.2: Exemple d'annotation des événements et des expressions temporelles, au format TimeML (XML). Une représentation graphique plus lisible est présentée à la figure 1.1.

maticales et aspectuelles de ces événements.

2. Identification et normalisation des expressions temporelles (balises `TIMEX3` dans le standard TimeML), en particulier les dates et les durées. La normalisation, pour une date, consiste à interpréter l'expression d'une date, souvent relative (par exemple, "`Le 24 juin dernier`" ou "`hier`"), pour en déduire la valeur absolue sous un format standard ("`2013-06-24`").
3. Extraction des relations temporelles entre :
   - les événements et la date de création du document ;
   - les événements et les expressions temporelles ;
   - les événements entre eux,

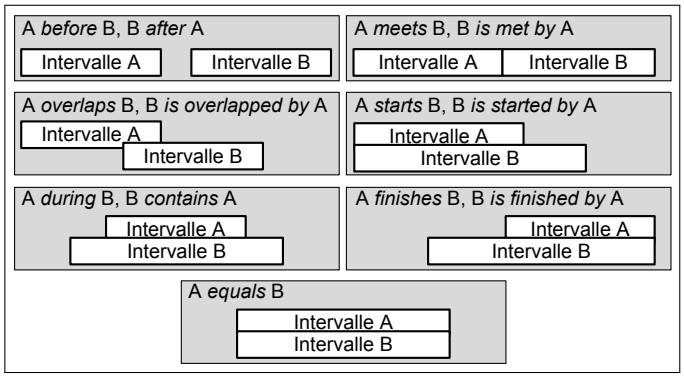

FIGURE 1.3: Les 13 relations d'Allen

conduisant à un ordre partiel entre les événements. Ces relations temporelles peuvent être choisies dans l'ensemble des 13 relations d'Allen [6] (voir figure 1.3) ou, une telle précision étant difficilement compatible avec l'interprétation de textes en langage naturel, dans un sous-ensemble de ces relations, pouvant aller jusqu'au simple trio "*before*", "*after*" et "*overlaps*" des campagnes TempEval [156, 157].

Ces stratégies se sont appliquées principalement à des textes du domaine général issus de corpus médiatiques, les travaux étant structurés et animés par le groupe de travail autour de la norme TimeML et par les trois campagnes d'évaluation TempEval en 2007 [156], 2010 [157] et 2013 [152].

Plus récemment, le domaine médical s'est emparé de la question, notamment au travers de la campagne d'évaluation i2b2 [144], avec une réflexion centrée sur de l'extraction de la chronologie des événements dans des comptes rendus d'hospitalisation. De tels comptes-rendus spécifient la date d'hospitalisation et la date de sortie de l'hôpital, références temporelles qui remplacent la fameuse "DCT" (date de création du document, ou *document creation time* en anglais) du domaine général. On y voit ensuite, structurées de façon plus ou moins explicite, des sections relatant l'histoire de la maladie, c'est-à-dire les événements ayant conduit à l'hospitalisation actuelle (prédispositions, maladies et hospitalisations précédentes, traitement, etc.), ainsi que l'évolution durant le séjour du patient dans les services de l'hôpital (nouveaux traitements, opérations chirurgicales, transferts, etc.). La figure 1.4 illustre un document annoté en anglais (les travaux sur le français n'en étant qu'à leurs débuts).

L'étude des relations temporelles entre événements dans les comptes-rendus d'hospitalisation a pour objectif de construire une chronologie des actions effectuées et des problèmes rencontrés, dans le but d'apporter une meilleure compréhension de la progression d'une maladie, de l'efficacité d'un traitement, ou de construire des systèmes d'aide à la prise de décision [69, 166, 137]. On peut également imaginer des applications en fouille de texte ; par exemple, en comparant les soins reçus et un protocole clinique type devant s'appliquer selon le profil du patient, on pourrait constater à quel moment on dévie de ce protocole (doses de médicaments différentes, moments différents) et analyser les raisons de ces divergences pour découvrir des intéractions nouvelles entre médicaments, des toxicités inconnues ou des impacts inattendus d'antécédents médicaux.

Ce qui est particulièrement intéressant lorsque l'on compare cette même tâche d'extraction de relations temporelles dans des textes du domaine général d'une part, et du domaine bio-médical d'autre part, c'est que l'on constate que la façon de décider ce

ADMISSION DATE : <TIMEX3 type="date" id="t1">**10/17/95**</TIMEX3>
DISCHARGE DATE : <TIMEX3 type="date" id="t2">**10/20/95**</TIMEX3>
HISTORY OF PRESENT ILLNESS :
This is a 73-year-old man with <EVENT type="problem" id="e1">squamous cell carcinoma of the lung</EVENT>, status post <EVENT type="treatment" id="e2">lobectomy</EVENT> and <EVENT type="treatment" id="e3">resection</EVENT> of <EVENT type="problem" id="e4">left cervical recurrence</EVENT>, <EVENT type="occurrence" id="e5">admitted</EVENT> here with <EVENT type="problem" id="e6">fever</EVENT> and <EVENT type="problem" id="e7">neutropenia</EVENT>.
(...)

HOSPITAL COURSE :
He was started on <EVENT type="treatment" id="e8">Neupogen</EVENT>, 400 mcg. subq. q.d.
(...)
He was <EVENT type="occurrence" id="e9">discharged</EVENT> home on <EVENT type="treatment" id="e10">Neupogen</EVENT>.

e3 OVERLAP e4	e1 OVERLAP t1	e2 BEFORE t1	e3 BEFORE t1
e4 BEFORE t1	e6 OVERLAP t1	e7 OVERLAP t1	e8 BEFORE t2
e9 OVERLAP t2	e10 OVERLAP t2		

FIGURE 1.4: Extrait d'article (anonymisé) du corpus i2b2/VA, montrant les événements (EVENT), les expressions temporelles (TIMEX3) ainsi que quelques relations temporelles au format simplifié.

qu'est un événement varie sensiblement de l'un à l'autre, c'est-à-dire qu'on ne fait pas porter "ce qui se passe" par les mêmes éléments de la phrase. Dans la norme TimeML et les campagnes TempEval du domaine général, on associe plutôt l'événement à la notion de *prédicat*, en annotant les verbes d'action de façon systématique, puis les noms ou adjectifs prédicatifs. Dans i2b2 et le domaine bio-médical, l'événement est essentiellement un groupe nominal relié à un *concept* [63]. Ainsi, parmi ces concepts/événements, on distingue les "occurrences", "problèmes médicaux", les "traitements", les "examens" et même les "services hospitaliers", mentionnés le plus souvent lors d'un transfert, donc d'un événement. Par exemple, pour la phrase suivante :

(1.1) `Le patient a été placé sous Neupogen.`

un adepte de TimeML décidera que "`placé`" est l'événement, et que le médicament (Neupogen) est l'objet cible, tandis que son confrère participant à i2b2 annotera plutôt "`Neupogen`", qui est finalement bien l'essentiel à retenir du changement d'état décrit par la phrase. Pourtant, aucun conflit de définition entre théoriciens ne se cache derrière cette différence : comme l'admettent volontiers les créateurs du corpus et de la norme bio-médicale [63], ce sont uniquement des considérations pratiques qui les ont conduits à se distinguer de cette façon. Parmi ces raisons, la nécessité de réutiliser des corpus existants, déjà annotés en concepts, ainsi que l'intérêt de pouvoir relier les "événements" aux terminologies du domaine, comme MeSH ou l'UMLS, qui contiennent des concepts et les organisent entre eux. Ainsi, annoter "`Neupogen`" dans l'exemple précédent permet de relier la mention de ce médicament à la base de connaissance, qui nous conduira par exemple aux maladies ou effets indésirables associés.

Que ce soit du point de vue de la représentation ou de celui des applications envisagées, la vision de l'événement décrite dans cette section souffre d'un certain nombre de limitations. Le choix de considérer l'événement comme un simple mot, sans prendre ses arguments en considération, en est une première, même si on peut l'utiliser comme un

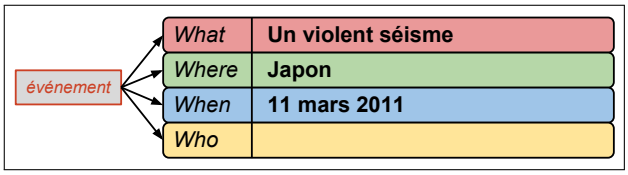

FIGURE 1.5: Représentation de l'exemple 1.2 sous le principe générique des 4W.

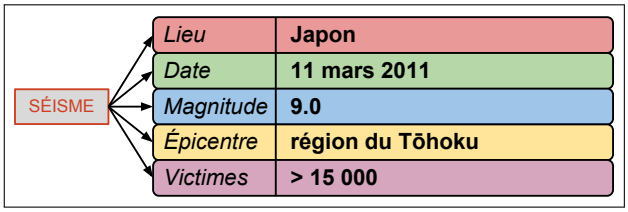

FIGURE 1.6: Représentation de l'exemple 1.2 avec un schéma spécifique prédéfini.

premier pas avant d'autres traitements. D'autre part, il s'agit d'une tâche très difficile (autour de 35 % de précision et de rappel lors de TempEval 3 pour la tâche complète) et pour laquelle on peut parfois douter, pour les textes du domaine général en tout cas, que les efforts faits pour améliorer les résultats en valent vraiment la peine. En effet, les relations les plus complexes à extraire sont celles qui sont implicites dans le texte, et elles ont finalement rarement un intérêt déterminant pour la compréhension du texte. Enfin, la tâche s'inscrit pleinement et uniquement dans le texte ; la norme TimeML est un balisage du texte, et les informations annotées ne peuvent s'extraire de ce texte. Par son attachement exclusif à la réalisation textuelle, cette représentation de l'événement ne peut s'étendre au-delà du document.

1.2 Un schéma informationnel

Une autre vision, plus proche de l'extraction d'information dite traditionnelle, se traduit par une approche relationnelle de la notion d'événement. "Ce qui se passe" est alors précisé par différents rôles joués par les entités gravitant autour de l'événement. Les "participants" (au sens parfois très large) de l'événement sont extraits du texte pour composer une sorte de fiche d'identité de cet événement.

Là encore, différents modes de représentation peuvent être choisis selon la granularité et la spécialisation visées. Par exemple, pour le texte suivant :

(1.2) Un violent séisme a secoué le Japon le 11 mars 2011. D'une magnitude de 9.0 en son épicentre dans la région du Tōhoku, il a officiellement fait plus de 15 000 victimes.

une extraction générique d'événements, basée sur les "4 W" factuels, chers aux journalistes (*what*, *where*, *when*, *who*), conduira à une structure telle que celle présentée à la figure 1.5. Mais la vision traditionnelle de l'extraction d'information, issue des célèbres campagnes d'évaluation MUC, mères de nombreux travaux dans le domaine [65], consiste à proposer un schéma spécifique à chaque type d'événement (comme par exemple, un séisme) en étudiant les caractéristiques qu'il est important de connaître à son sujet (pour un séisme, en plus de sa date et de son lieu, sa magnitude, son épicentre, le nombre de victimes). La figure 1.6 illustre ce type de schéma.

Notons que si cette approche relationnelle n'est pas spécifique à l'événement, les travaux concernant l'événement ont pris une grande importance dans cette communauté.

Une fois que les attributs d'un type d'événement sont fixés, il est possible de remplir des tables de données concernant des événements de ce même type, avec des informations provenant de plusieurs sources. Ceci conduit certes à un élargissement de l'horizon de recherche et à de nouveaux potentiels applicatifs, mais pose également de nouveaux problèmes, comme l'homogéisation des données multidocuments, la normalisation des informations, la gestion des conflits. Par exemple, si un document mentionne le séisme du "Japon", et un autre, à la même date, celui de "l'île de Honshū", comment savoir qu'il s'agit du même événement ? Si un document mentionne "plus de 15 000 victimes", un autre "15 776", et un troisième "21 000", comment remplir le schéma ?

D'autre part, l'inconvénient de cette représentation spécifique est qu'il semble nécessaire de connaître à l'avance les spécificités d'un type d'événement particulier pour construire une structure informationnelle appropriée, mais aussi pour la remplir. On a alors un recours coûteux à des experts du domaine, que ce soit pour annoter manuellement des quantités importantes de textes (dans le but d'entraîner des systèmes d'apprentissage), pour alimenter des systèmes à base de règles, ou pour construire des ontologies spécialisées complexes. Un système d'extraction d'information reste alors limité à un domaine spécifique, l'adaptation à un nouveau domaine impliquant de nouveaux coûts.

Cependant, depuis que le traitement de très grands volumes de textes est envisageable à un coût raisonnable, le fait de s'élever au-dessus du texte et d'en extraire des composants offre de nouvelles perspectives. On peut ainsi dépasser ce seul angle des conférences MUC et de la tâche consistant à extraire de textes des entités courtes et à leur attribuer un rôle dans une structure informationnelle définie à l'avance.

En effet, profusion d'information rime le plus souvent avec redondance d'information. Si cela constitue parfois un problème à résoudre (renvoyer plusieurs fois la même information à l'utilisateur n'est pas pertinent), la redondance peut souvent s'avérer précieuse. Grâce à elle, on peut notamment :

- compléter des informations sur un événement donné, si aucun document ne fournit la totalité des renseignements ;
- confirmer des informations et leur attribuer éventuellement une valeur de confiance (si l'on rencontre une même information de nombreuses fois, celle-ci devient plus sûre) ;
- estimer l'importance des événements, en vertu du principe (certes discutable) qu'un événement est important si l'on en parle beaucoup ;
- étudier l'évolution dans le temps des données concernant un événement (changement de point de vue, changement de mode de désignation, évolution des données numériques associées, réactions...).

Toutes ces opportunités ont conduit les chercheurs en extraction d'information en général, et ceux travaillant sur les événements en particulier, à imaginer des tâches plus diverses et moins contraintes, permettant de s'adapter à des problèmes nouveaux et à des contextes dans lesquels le besoin informationnel est de moins en moins stable ou de moins en moins bien défini.

Dans un premier temps, on utilise la quantité de texte présente dans les grandes collections pour réduire le coût lié au recours à l'expertise humaine, tout en restant cantonné à des systèmes n'extrayant l'information que pour un domaine particulier. On souhaite alors limiter le temps consacré soit à l'annotation manuelle des textes de référence, soit à l'élaboration de règles d'extraction. Les méthodes employées visent à se contenter de quelques règles ou de quelques exemples, en employant des méthodes hy-

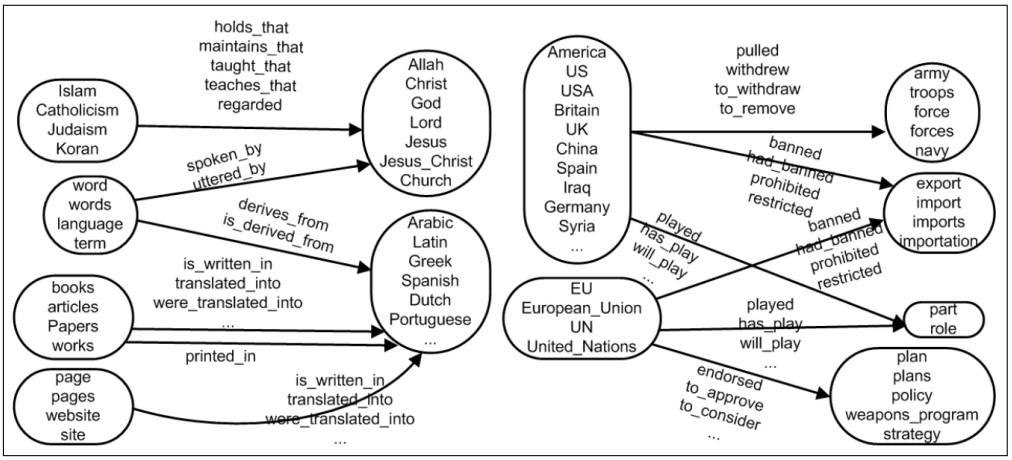

FIGURE 1.7: Exemple de réseau sémantique induit automatiquement [92].

brides [60], de l'apprentissage actif (*active learning* – [139]), des mécanismes d'amorçage (*bootstrapping* – [73, 1]) ou de la "supervision distante" [116]. On parle alors d'approches faiblement supervisées, favorisées dans certains cas par la mise à disposition de larges bases de connaissances (comme DBPedia [98] ou OpenCalais [1]).

L'apport de ces techniques est une plus grande souplesse et une réduction du coût dans la construction du système, mais pas une fonctionnalité ni un mode de représentation nouveaux. Une autre piste consiste alors à supprimer totalement ou presque la phase de supervision, et à éviter de définir *a priori* les types de structures et de relations recherchés. On utilise la grande quantité d'information disponible pour établir des regroupements et découvrir automatiquement les structures et des relations, par exemple en mettant des motifs en évidence dans un ensemble de documents du même thème [138, 141].

Tous ces travaux conduisent de façon relativement naturelle à une application encore plus centrée sur l'exploitation de la masse de données disponibles, s'éloignant encore plus du texte et conduisant à un mode de représentation encore différent. On extrait en effet des relations entre entités de façon non supervisée dans le but d'acquérir des connaissances du monde, non pas sur des événements particuliers mais sur des relations génériques qu'entretiennent les entités ou les événements entre eux.

On obtient ainsi des graphes représentant des connaissances générales sur les événements courants et leurs enchaînements habituels. Dans le cas de l'extraction de "réseaux sémantiques" [15, 92], les relations extraites sont des événements qui relient des groupes d'entités (nœuds du graphe), comme l'illustre la figure 1.7. On y voit par exemple qu'un "livre", un "article", un "papier" ou un "travail" peuvent être "*traduits en*" "arabe", "latin", "grec", etc.

Dans les "chaînes d'événements" [33], en revanche, les événements sont les nœuds du graphe, et non les relations, celles-ci indiquant la précédence entre deux événements (relation *before*). On obtient ainsi, comme le montre la figure 1.8, des enchaînements ou des procédures typiques, dont une des applications peut être l'aide à la détermination des relations temporelles, telle que décrite à la section précédente.

Ces approches, qui mettent en œuvre à la fois des traitements linguistiques fins et une abstraction statistique, peuvent être menées sur des textes du domaine général,

1. http://www.opencalais.com

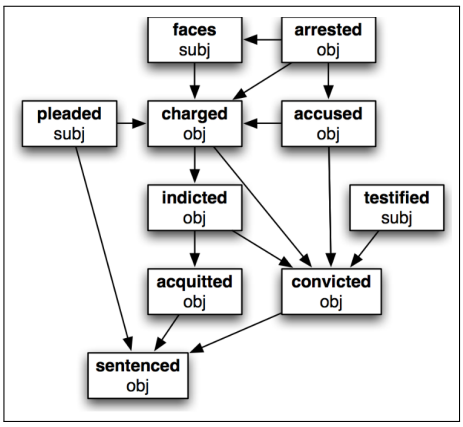

FIGURE 1.8: Exemple de chaîne d'événements sur le thème des poursuites judiciaires, induite automatiquement [33]. Les flèches indiquent une relation de précédence.

mais aussi sur un domaine de spécialité, comme le domaine biomédical, où les résultats peuvent ensuite être intégrés à une ontologie existante [123, 38].

Ainsi, la représentation relationnelle de l'événement permet d'aggréger des informations en provenance de nombreux documents, d'extraire des informations très précises sur les caractéristiques des choses qui se passent et de construire des bases de connaissances qui ont prouvé leur utilité. Elle ne traite malgré tout qu'une certaine catégorie de problèmes et de besoins d'information : ceux qui peuvent être satisfaits par un ciblage d'information vers des éléments courts, précis et factuels. D'autre part, cette approche est de loin la plus ancienne et la plus étudiée, et une amélioration des performances nécessite maintenant des progrès sensibles et combinés dans le traitement de tous les phénomènes de la langue qui perturbent traditionnellement les analyses : variations lexicales, syntaxiques, sémantiques, divers problèmes de coréférence, normalisation d'entités nommées, etc.

1.3 Une description textuelle complexe

Une vision radicalement différente, même si elle est souvent utilisée en préalable à la précédente, conduit à se contenter de considérer l'événement tel qu'il a été décrit par l'auteur d'un document, sans chercher à le réduire ni à le structurer. Autrement dit, un événement est identifié à la portion de texte qui le décrit, qu'il s'agisse d'une partie de phrase, d'une phrase complète, d'un ensemble de phrases ou d'un document entier.

Si les chercheurs nomment toujours ce qu'ils font de la même façon (*event extraction*, *event detection*) alors que les notions manipulées, les tâches et les objectifs sont clairement différents des stratégies présentées précédemment, le vocabulaire employé par ailleurs emprunte plus au domaine de la recherche d'information que de l'extraction d'information : on parle de segmentation d'événements, de similarité, d'ordonnancement, de regroupement (*clustering*). Le mode de représentation le plus fréquent est la représentation vectorielle, l'événement devient donc un simple sac de mots dans lequel l'ordre des mots dans les phrases est perdu (figure 1.9).

Cette approche rejoint pourtant parfois celle du schéma informationnel, proposant des relations entre entités, décrit à la section précédente. Ainsi, plutôt que de voir la relation entre deux entités comme un prédicat abstrait, étiqueté par un simple mot (figure 1.7), on peut la caractériser dans le texte, par les éléments du contexte phrastique

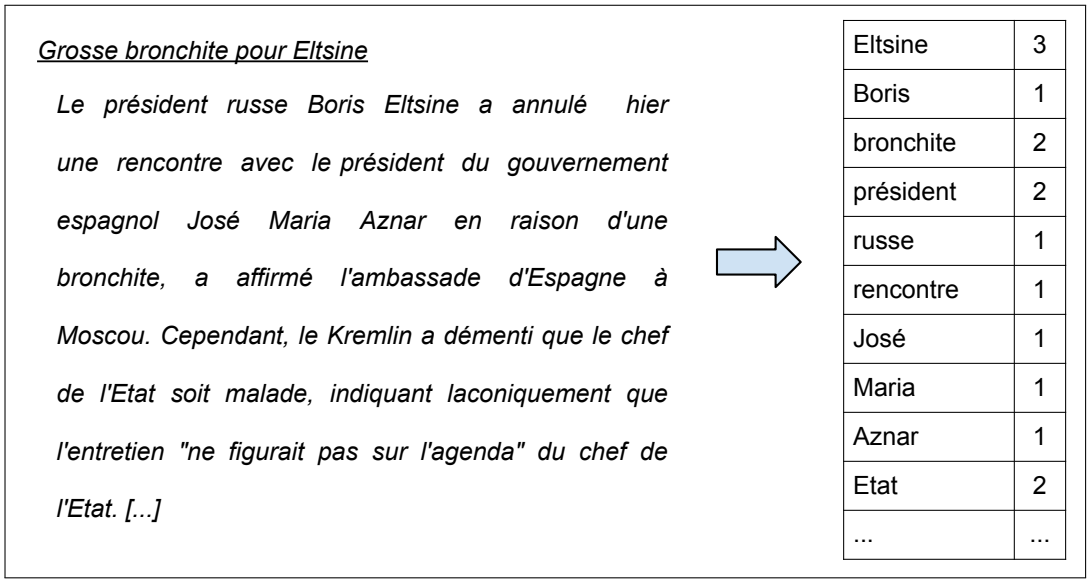

FIGURE 1.9: Représentation vectorielle d'un texte. Les mots ou termes peuvent être associés à des poids, ici le nombre d'occurrences dans le document.

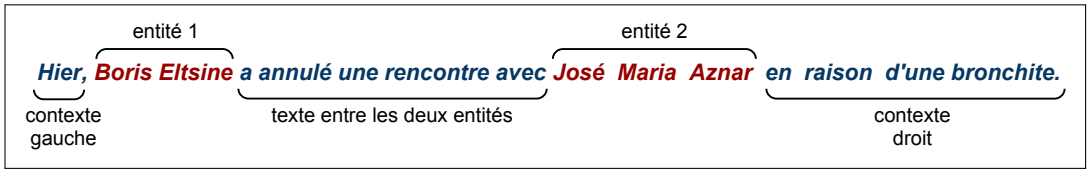

FIGURE 1.10: Représentation d'un événement par une relation entre deux entités, caractérisée par le texte situé entre les deux entités, le contexte gauche et le contexte droit [158].

de ces entités [158] : la partie de texte entre les entités, la partie de texte située avant la première entité et la partie située après la seconde entité (voir exemple figure 1.10). On pourrait y voir un retour à l'attachement de la notion d'événement à la réalisation linguistique textuelle, et donc à la vision mono-document décrite dans la section 1.1 ; il ne s'agit au contraire que d'un premier pas conduisant, par l'intermédiaire de la représentation vectorielle, à des regroupements de relations provenant de nombreux documents (*clustering*).

Mais les applications les plus étudiées en la matière s'éloignent totalement de tout cela. Elles consistent d'une part en l'agrégation d'articles couvrant le même événement [111], et d'autre part en la création automatique de résumés multidocuments [112]. Ces résumés, généralement basés sur des corpus d'articles de presse, sont presque systématiquement centrés sur la notion d'événement [111], jusqu'à être organisés explicitement comme une chronologie d'événements importants [87] (figure 1.11). De telles applications ont pour but de présenter à l'utilisateur une vision synthétique de l'actualité dans une certaine période de temps (agrégation de nouvelles) ou sur un certain thème (résumé).

Les frontières qui séparent les descriptions d'événement dans les textes ne sont pas explicites, et sont même souvent inexistantes : une même phrase peut mentionner plusieurs événements, revenir à un événement mentionné auparavant, ou décrire

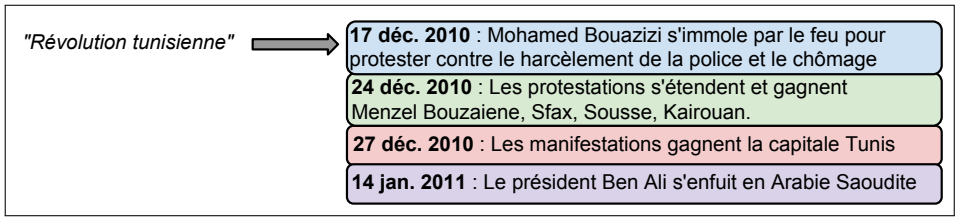

FIGURE 1.11: Élaboration d'une chronologie à partir d'une requête d'utilisateur. Un événement est ici représenté par une description textuelle associée à une date.

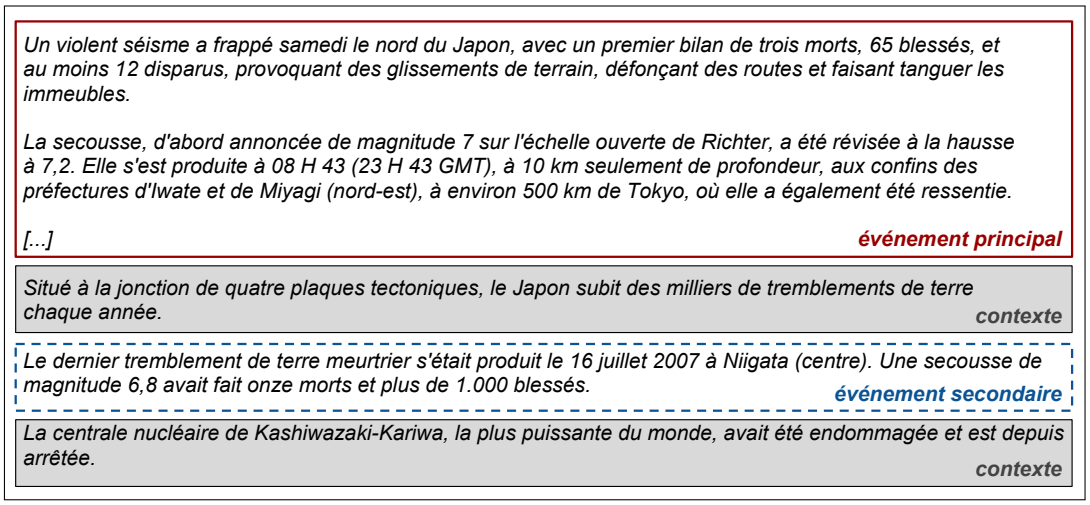

FIGURE 1.12: Segmentation des descriptions textuelles d'événements dans un texte, selon [78].

des "sous-événements" constitutifs d'un autre. C'est pourquoi cette approche conduit nécessairement à des approximations et du bruit. Elle permet cependant de ne pas se limiter à un mot ou à quelques attributs, et de tenir compte du contexte et de la séquence des descriptions, étant entendu que des techniques statistiques devront ensuite se charger de gérer ce bruit par filtrage ou agrégation. Elle nécessite également souvent une classification préalable des unités textuelles considérées. Ainsi, si l'on considère le document dans son ensemble, on peut distinguer les documents centrés sur un événement particulier (par exemple, l'éruption d'un volcan), sur une personne (avec des informations d'ordre biographique), sur plusieurs événements (événements du même ordre, repris dans le même article) [110, 104]. Si l'on considère une unité plus réduite à l'intérieur d'un document (une phrase ou une séquence de phrases), on cherchera à distinguer l'événement principal des événements secondaires [89], ou à opérer une classification plus précise encore : contexte [78], événements antérieurs, conséquence de l'événement principal [39], etc. Un exemple, issu de [78], est proposé à la figure 1.12.

La sortie des systèmes d'agrégation ou de résumé est alors, comme l'entrée, du texte brut, accompagné ou pas d'informations supplémentaires (date et lieu de l'événement concerné, source). Dans le cas de l'agrégation, le texte original n'est pas modifié mais simplement organisé de façon synthétique, comme on le voit par exemple dans les agrégateurs de nouvelles sur le Web (figure 1.13). Dans le cas du résumé, on peut soit composer un résumé à partir de phrases sélectionnées dans les documents originaux, par exemple à l'aide de mesures de représentativité ou de centralité dans des *clusters* de phrases similaires établis au préalable [64] ; soit produire de nouvelles phrases,

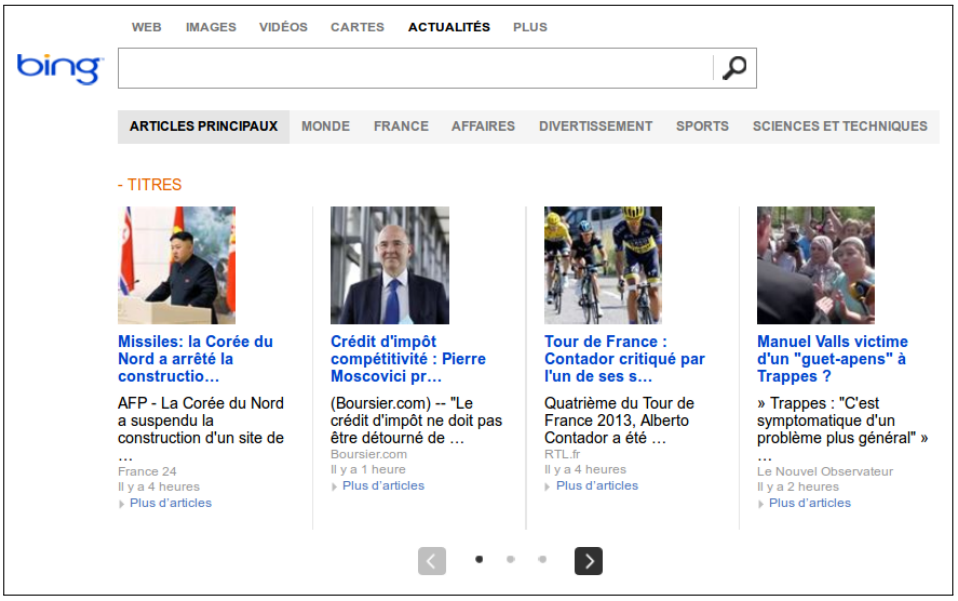

FIGURE 1.13: Un agrégateur de nouvelles (ici, celui du moteur Bing) regroupe les articles par événements, en choisit un représentatif et propose des liens vers l'ensemble des autres articles.

condensé des informations jugées importantes dans ces mêmes *clusters*, on parle alors de fusion [19] ou de compression [56]. Dans le premier cas, on dépend de la qualité et du caractère informatif et synthétique des phrases disponibles, et on risque de présenter des phrases avec des anaphores non résolues, si l'antécédent apparaît dans une phrase non sélectionnée. Dans le second cas, on maîtrise mieux la génération du résumé, mais on ajoute l'inconvénient de produire parfois des phrases syntaxiquement incorrectes.

On voit qu'une nouvelle dimension apparaît ici dans la représentation de l'événement, à savoir l'importance de cet événement. Dans un résumé ou une chronologie automatiques, qu'ils soient basées sur une requête d'utilisateur ou sur un ensemble de documents à résumer, seuls les événements les plus importants devront apparaître.

1.4 Un pic d'activité thématique

Cet intérêt envers l'importance d'un événement est naturel dès que l'on se fixe pour objectif d'aider l'utilisateur à s'informer ou à apprendre, en extrayant pour lui la substantifique moëlle du déluge d'information qui s'abat sur lui quotidiennement.

Cette importance dépend bien entendu des thématiques qui intéressent l'utilisateur, généralement représentées par des mots-clés issus d'une requête explicite ou d'un profil construit semi-automatiquement. Elle dépend également de l'intérêt intrinsèque de l'événement, plus difficile à définir. Faute de mieux, celui-ci est généralement mesuré par la redondance de la désignation de l'événement dans les documents, ainsi que parfois par des critères temporels [87].

Si l'on poursuit cette piste de la caractérisation de l'événement par son importance (ou sa *saillance*), la représentation de celui-ci devient alors purement statistique : un événement n'est plus un segment textuel, ni une structure contenant du texte, mais un pic d'activité associé à une certaine thématique.

C'est la vision de deux grandes classes d'applications :

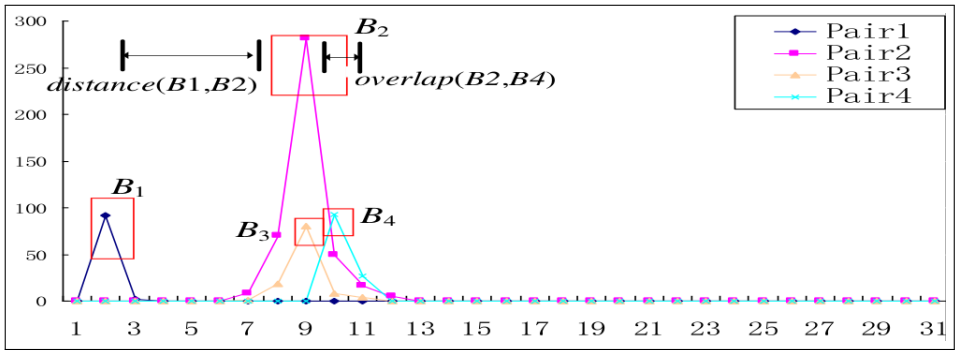

FIGURE 1.14: L'événement selon [67], un pic d'activité à distinguer du flux continu d'information et à comparer avec les autres pics.

- La "détection rétrospective d'événements" (RED pour *Retrospective Event Detection* [161]), dont le but est de détecter, dans un ensemble fixe de documents, soit tous les événements, soit tous les événements saillants, pour une thématique donnée.
- Et surtout, la "détection d'événements nouveaux" (NED pour *New Event Detection*), telle que définie en particulier par les campagnes TDT (*Topic Detection and Tracking* [5]). Dans cette tâche, on considère un corpus dynamique de nouvelles, avec un flux de nouveaux documents arrivant en permanence. Le but est alors de prédire si un nouveau document traite ou pas d'un nouvel événement (c'est-à-dire, un événement non encore abordé dans les documents précédents).

Si les documents en question ont longtemps été des articles de presse ou des dépêches d'agence, ces dernières années ont vu traiter de nouveaux textes comme des pages Web, des logs de recherche [67] et des tweets [20], conduisant à des quantités de documents bien plus importantes et dynamiques, ainsi qu'à de nouveaux types de problèmes : orthographe (syntaxe défaillante, nombreuses abréviations...) mais aussi textes plus courts avec moins de contexte et redondance très largement accrue, avec certains sous-thèmes et événements noyés dans un flot d'événements sur-représentés.

Dans la résolution de ces tâches, on considère plus que jamais les textes comme des sacs de mots, et on voit apparaître le vocabulaire et les techniques statistiques de la détection d'événements au sens d' "anomalies", en provenance des domaines de l'analyse des séries temporelles ou des données issues de réseaux de capteurs, pour lesquels une anomalie est un pic inhabituel d'événements relevés. Ici, ce ne sont pas des capteurs surveillant des flux d'eau, d'air ou de gaz, mais des analyses repérant des mots et surveillant des fréquences inhabituelles de certains mots ou combinaisons de mots (figures 1.14 et 1.15).

1.5 Conclusion et perspectives

Voilà donc l'événement décrypté dans tous ses états. Si nous avons déjà annoncé que nous ne chercherions pas à définir la notion d'événement de façon plus précise que "quelque chose qui se passe", il semble tout de même intéressant de réfléchir, durant quelques lignes, aux différents enjeux que cet effort représenterait. Nous pouvons attendre d'une définition de l'événement qu'elle détermine un ensemble de propriétés essentielles pour cette notion, dans le but de permettre de distinguer ce qui est un événement de ce qui n'en est pas un. Selon les différentes visions énumérées dans ce

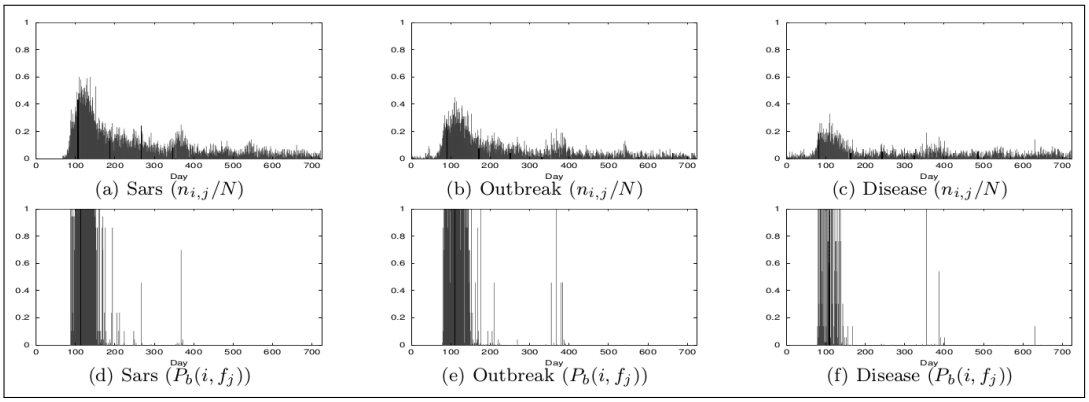

FIGURE 1.15: L'événement selon [62], une évolution de la saillance d'un thème au fil du temps (première ligne pour 3 termes, le pourcentage d'articles mentionnant le terme), à traduire mathématiquement par une probabilité P d'être un événement saillant (seconde ligne pour les mêmes termes).

chapitre, certaines de ces caractéristiques peuvent être :

- l'ancrage spatio-temporel ;
- la structure ;
- l'importance ;
- la nouveauté ;
- les relations entretenues avec les autres événements.

Ces propriétés doivent être distinguées d'attributs couramment manipulées lorsque l'on traite des événements, comme la factualité (le niveau de certitude quant au fait que l'événement a bien eu lieu), la temporalité (passé, futur), la récurrence (événement unique, multiple, périodique). Ces dernières ne sont pas des propriétés définitoires, mais plutôt des éléments de catégorisation, ou de typologie, des événements.

On voit donc sans surprise que la définition de l'événement dépend du point de vue et des tâches que l'on se fixe. On peut regrouper ces tâches selon les axes suivants :

- **Ciblage** : en lien avec le domaine de l'extraction d'information, il s'agit de pointer dans les textes les informations pertinentes sur les événements décrits, et éventuellement de dégager une structure et des relations entre ces informations.
- **Agrégation** : en ajoutant l'aspect multidocument à la tâche de ciblage, on agrège des propriétés glanées dans une collection de documents, pour obtenir une vision cohérente et structurée d'un même événement.
- **Hiérarchisation** : ici, l'importance des événements entre en jeu, et il s'agit de proposer à un utilisateur une vision résumée des événements principaux concernant un thème ou une période de temps.
- **Contextualisation** : lorsqu'un nouvel événement se produit, le but est de le détecter et d'apporter au lecteur les éléments essentiels pour le comprendre et le mettre en perspective.

Le tableau 1.1 résume les liens entre ces axes et les différents modes de représentation décrits dans ce chapitre. Les derniers axes sont associés à une représentation de moins en moins structurée. Cette compartimentation est d'ailleurs une limite importante des approches actuelles. On peut prédire que des travaux futurs auront pour but, ou au moins pour effet, de compléter la matrice formée par cette figure et de créer de nouveaux liens entre axes applicatifs et représentation.

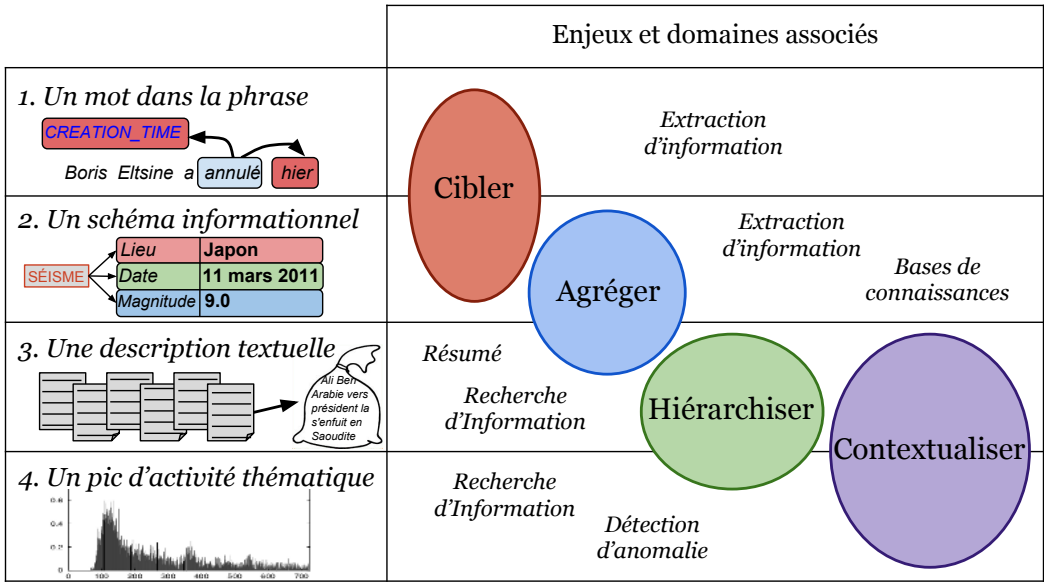

TABLE 1.1: Traitement des événements : enjeux et domaines associés.

Par exemple, lorsqu'on travaille sur l'extraction des désignations d'événement dans les textes et sur l'induction des relations temporelles entre ces événements (section 1.1), on constate vite que la complexité de la tâche vient en partie de la difficulté de choisir les paires d'événements qu'il est pertinent de lier temporellement. En effet, si des relations temporelles existent entre tous les événements, que ce soit dans un texte ou au-delà, seules certaines d'entre elles sont importantes pour la compréhension d'un texte. Ici, une bonne *hiérarchisation* de l'information dans le document permettrait d'améliorer les systèmes.

Dans l'autre sens, une fois que l'on a détecté des occurrences statistiques caractéristiques d'un événement (pic thématique par exemple), un problème qui se pose est de retransmettre cette information à l'utilisateur en recomposant une information textuelle sous forme de phrases (le texte qui décrit l'événement détecté). Autrement dit, une fois la hiérarchisation faite à l'aide de critères essentiellement statistiques, il est difficile de *cibler* de nouveau vers une description unique, concise et non redondante.

Quelques-uns des travaux décrits dans les chapitres suivants visent à combler ces lacunes et à apporter des éclairages originaux sur l'analyse des événements.

À ces quatre axes, nous souhaitons enfin ajouter celui de l'**évolution**, très peu exploré jusqu'à présent. Cette notion peut englober bon nombre d'aspects différents :

- Nouvelles informations sur un événement précis – par exemple, un enlèvement est revendiqué quelques jours après qu'il a eu lieu.
- Évolution de données numériques concernant un événement – typiquement, l'affinement d'un bilan chiffré.
- Évolution de l'importance accordée à un événement – par exemple, l'immolation de Mohamed Bouazizi est anodine jusqu'à ce que l'on considère qu'elle a déclenché la révolution tunisienne.
- Évolution de la façon de désigner un événement – de *"deux avions se sont écrasés sur les tours du World Trade Center à New York"* jusqu'à *"11-Septembre"*.

Nous pouvons également ajouter à cette liste l'étude des relations entre événements (cause, conséquence, réaction), plus présente dans la littérature. Cet axe peut enfin être relié aux réflexions sur la plage temporelle de validité des schémas collectés en extraction d'information générale – François Hollande est premier secrétaire du P.S., député de Corrèze, président de la République, à des dates différentes.

Les progrès récents de l'analyse temporelle des textes, d'une part, et les liens de plus en plus étroits entre une frange du traitement automatique des langues et la fouille de données, d'autre part, permettent de laisser penser que des travaux sur l'évolution des événements, sous les angles décrits ci-dessus, devraient émerger et conduire à de nombreuses applications innovantes.

Forts des constats et des réflexions de ce chapitre, les deux chapitres suivants s'appuient sur les axes décrits ici, mais sont organisés selon des critères différents, pour apporter un éclairage supplémentaire. Nous distinguons ainsi "l'événement dans le texte", c'est-à-dire la désignation linguistique de quelque chose qui se passe, et "l'événement dans le monde", la perception de ce qui se passe par celui qui en est informé.

Chapitre 2

L'événement dans le texte

Extraction de désignations d'événements et relations temporelles

Explorer la question de l'événement dans un document textuel revient à se poser deux questions :

- Quels sont les événements décrits dans le texte ? (et donc, en ce qui nous concerne, quels sont les mots ou les groupes de mots qui désignent des événements dans le texte ?)
- Quelles sont les relations qui lient ces événements ?

Les types de relations possibles sont nombreux (cause/conséquence, contraste, élaboration, etc.) et décrites dans plusieurs théories comme la RST (Rethorical Structured Theory [108]) ou la SDRT (Segmented Discourse Representation Theory [95]). Dans notre cas, nous nous intéressons uniquement aux relations purement temporelles. Ces relations temporelles peuvent être choisies dans l'ensemble des 13 relations d'Allen [6] (voir figure 2.1) ou, une telle précision étant difficilement compatible avec l'interprétation de textes en langage naturel, dans un sous-ensemble de ces relations, pouvant aller jusqu'au simple trio "*before*", "*after*" et "*overlaps*" des campagnes d'évaluation TempEval [156, 157].

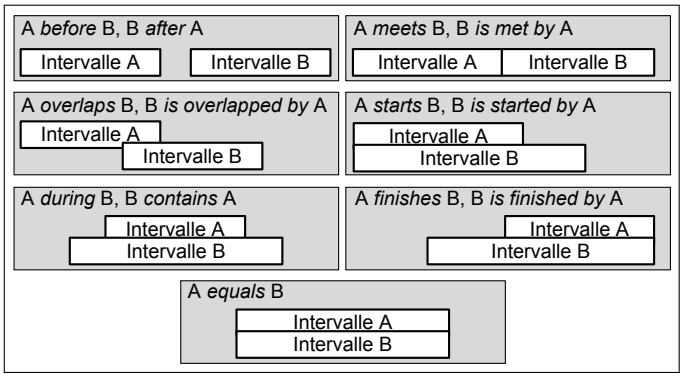

FIGURE 2.1: Les 13 relations d'Allen

L'importance de la composante temporelle en extraction d'information a conduit au développement de systèmes d'annotation de l'information temporelle, en grande majorité pour l'anglais. Plusieurs de ces systèmes sont présentés et évalués dans le cadre des campagnes TempEval [156, 157, 152]. Ils sont dans la plupart des cas basés sur l'apprentissage à partir de corpus annotés selon la norme ISO-TimeML. Alors qu'en 2007 [156], les tâches d'annotation principales étaient centrées sur la détermination de relations temporelles entre événements et/ou expressions temporelles pré-annotées, en 2010 [157] et 2013 [152], deux tâches consistent en la reconnaissance et le typage automatique des expressions temporelles et des événements sans annotation préalable. Les efforts de la communauté tendent ainsi vers une tâche d'annotation temporelle partant du texte brut et allant jusqu'à une analyse fine de la structure temporelle d'un document. Le travail de [151] est une illustration de cette préoccupation. Il présente un système combinant analyse syntaxique profonde et classifieurs pour procéder à l'annotation et la normalisation des événements et des expressions temporelles à partir de textes tout venant. La plupart des systèmes existants et ayant participé aux compétitions mentionnées ci-dessus traitent des textes en anglais, ce qui est notre cas également pour les travaux décrits ici [1].

Nous présentons tout d'abord nos travaux sur l'identification des désignations nominales dans les textes (section 2.1). Puis nous nous intéressons à l'identification de relations temporelles dans des textes du domaine général (section 2.2) et dans le domaine médical (section 2.3). Enfin, nous décrivons le problème de l'évaluation des graphes de relations temporelles et les métriques que nous avons proposées pour améliorer cette évaluation (section 2.4).

2.1 L'événement nominal

Un événement peut être évoqué par une description verbale ou nominale. Ainsi, on pourra mentionner des résultats d'élections dans une phrase dont le verbe porte l'événement (description verbale de l'événement) :

(2.1) Le président n'a pas renouvelé son mandat.

(2.2) l'UMP a perdu la bataille pour l'Élysée.

Ou dans un groupe nominal (description nominale de l'événement) :

(2.3) La défaite du président sortant [...]

(2.4) L'échec de l'UMP aux élections présidentielles [...]

De nombreux travaux en traitement automatique de langues et extraction d'information se sont focalisés sur la description verbale des événements, mais peu d'entre eux se sont intéressés aux désignations nominales des événements [23]. Il faut noter que lorsque l'événement n'est pas désigné sous une forme verbale, il a de grandes chances de l'être sous une forme nominale. Ne traiter que les verbes, c'est donc prendre le risque de manquer les informations recherchées.

Quant aux études linguistiques à ce sujet, elles s'attachent plutôt à la description détaillée d'un ou de quelques événements précis, comme la *révolution du Jasmin*, la *grippe H1N1* ou le *11-Septembre* [31], mais ne tentent jamais d'apporter un cadre de généralisation.

1. Les quelques travaux sur le français que nous avons menés [88, 12, 11] ne présentent pas de différences méthodologiques significatives avec ceux portant sur l'anglais, c'est pourquoi nous ne les abordons pas.

Nous nous plaçons pour notre part dans un cadre d'extraction d'information, ce qui nous pousse à considérer la désignation nominale d'événement comme une entité nommée [8] [1].

2.1.1 Modes de construction des noms d'événement

Le verbe, à l'exception du verbe d'état, a pour rôle principal de représenter des événements. Il est voué à cet usage, et renferme notamment des propriétés d'aspect, de mode, de temps [155] qui permettent de déterminer ces événements de façon plus spécifique.

Le nom en revanche est moins facile à cerner, en particulier par un outil d'extraction automatique, puisqu'il n'a ni cette vocation intrinsèque ni ces propriétés de surface. Nous avons identifié trois modes de construction des noms d'événement [13] :

1. Nominalisation apparentée à un verbe d'action, c'est-à-dire ce que nous pourrions appeler par abus de langage des déverbaux [2]. Tous ces noms n'ont pas une lecture événementielle et ont en général au moins une autre lecture possible, souvent celle du résultat de l'action ("`construction`", "`exil`"), de son instrument ("`shampooing`"), de sa localisation ("`étalage`", "`arrivée`"), etc.
2. Nominalisation autre que dérivée de verbes, avec des noms décrivant intrinsèquement des événements, comme "`festival`", "`match`", "`tsunami`", "`apartheid`". L'ambiguïté existant sur ces noms, moins importante, est celle liée à l'homonymie et à la polysémie "classiques" (comme pour les mots "`salon`" – l'événement ou la pièce d'un logement – ou "`triathlon`" – le sport ou l'épreuve).
3. Nominalisation par métonymie, pour des noms représentant des événements dans des contextes précis, comme les toponymes événementiels (noms de lieu désignant un événement [96], exemple 2.5), les héméronymes (dates désignant un événement [30], exemple 2.6) ou de simples noms d'objet (exemple 2.7).

   (2.5) `Personne ne veut d'un nouveau` *`Tchernobyl`*`.`

   (2.6) `On pourrait assister à un` *`21 avril`* `à l'envers.`

   (2.7) `Les` *`frégates de Taïwan`* `s'invitent à la conférence de Lorient.`

   Dans ce cas, l'ambiguïté est par nature systématique.

La distinction entre ces trois classes est importante en vue d'une extraction des noms d'événement. Pour la première classe, des indications morphosyntaxiques sur une conversion en lien avec un verbe seront utiles. Dans le deuxième cas, la constitution d'un lexique semble nécessaire. La troisième classe pose un problème plus rare, mais plus difficile.

2.1.2 Étude du phénomène en corpus

La désignation d'événements par des noms étant très peu étudiée de façon systématique, il nous a semblé intéressant de réaliser une étude quantitative de l'utilisation des noms d'événement et de leur ambiguïté dans les textes.

---

1. Les travaux décrits dans cette section ont été réalisés avec Béatrice Arnulphy et Anne Vilnat, avec le soutien financier du projet Quaero.
2. Nous nous intéressons aux paires nom-verbe apparentées car formées sur des bases identiques, même si la dérivation et son sens ne sont pas avérés [149].

Création du corpus

La création d'un corpus était de toute façon rendue nécessaire par la volonté d'entraîner et d'évaluer un système d'annotation automatique des événements nominaux. Nous avons élaboré un guide d'annotation suivant les préconisations générales des travaux sur les entités nommées du projet *Quaero* [10] [1]. Ce guide définit une typologie des événements, fournit de nombreux exemples, ainsi que des instructions permettant de décider si un nom ou un groupe nominal désigne un événement ou pas.

Nous avons ensuite annoté manuellement 192 articles de presse des journaux *Le Monde* et *L'Est Républicain*. Ce corpus de 47 646 mots au total comprend 1844 noms d'événements. À titre de comparaison, le corpus anglais TimeBank [126] en compte 1792, le corpus italien IT-TimeBank [134] 3695, le *FR-TimeBank* [24] 663.

Ce travail d'annotation et le corpus ont été présentés dans [14].

Comportement des noms en lecture événementielle

Nous avons ainsi pu observer le comportement des noms d'événements annotés, tout d'abord selon les trois modes de composition décrits ci-dessus. Ainsi, 67 % des événements nominaux du corpus appartiennent à la première catégorie (nominalisation d'un verbe d'action), 32 % à la deuxième (nominalisation sans qu'un verbe associé n'existe), et seulement 7 occurrences ont un caractère événementiel par métonymie.

Plus important, nous avons étudié l'ambiguïté de chacun de ces noms. Le corpus contient 725 occurrences différentes parmi les 1844 annotations. Parmi ces événements, 273 n'apparaissent qu'une seule fois. Sur l'ensemble des 452 noms d'événements annotés plus d'une fois comme événement, seuls 31 % ont une lecture événementielle à chacune de leur occurrence, les autres sont ambigus. La figure 2.2 donne plus de chiffres sur la proportion des lectures événementielles par rapport au nombre total d'occurrences d'un mot. On y remarque notamment que pour 29 noms du corpus, la proportion entre les occurrences où ils désignent un événement et le nombre total d'occurrences est de moins de 10 %. Pour 129 noms, ce rapport est inférieur à 50 %, et pour 312 noms, ce rapport est de moins de 100 % (soit au plus 99 % de leurs occurrences).

Quelques exemples en fonction du taux d'événementialité des mots extraits sont également présentés dans cette figure 2.2.

Pour plus de clarté, nous avons expliqué l'utilisation de ces mots au moyen d'exemples en contexte :

- "`Prix`", "`mort`", "`conseil`" ou "`triathlon`" sont à moins de 40 %.

  (2.8) `Dans ce <event>grand prix </event> au millimètre, seules neuf voitures ont passé la ligne d' <event> arrivée </event>.` (événement sportif)

  (2.9) `ce produit décal [...] est distribué depuis septembre au prix`$_{(montant)}$ `de 15,90 francs`

  (2.10) `Jusqu'à la <event> mort </event>.` (passage de la vie au trépas)

  (2.11) `Un mort`$_{(pers)}$ `dans un dépassement.`

- "`Commentaire`", "`signe`", "`prescription`" et "`bombe`" sont entre 40 et 69 %.

  (2.12) `À l'occasion du <event> 54ème anniversaire des <event> bombes </event> de Hiroshima et Nagasaki </event>` (par métonymie)

  (2.13) `Une bombe`$_{(obj)}$ `a explosé hier.`

- "`Campagne`" ou "`peine`" sont événement dans 70 à 99 % des occurrences.

1. Avec l'aide de Cyril Grouin et Sophie Rosset.

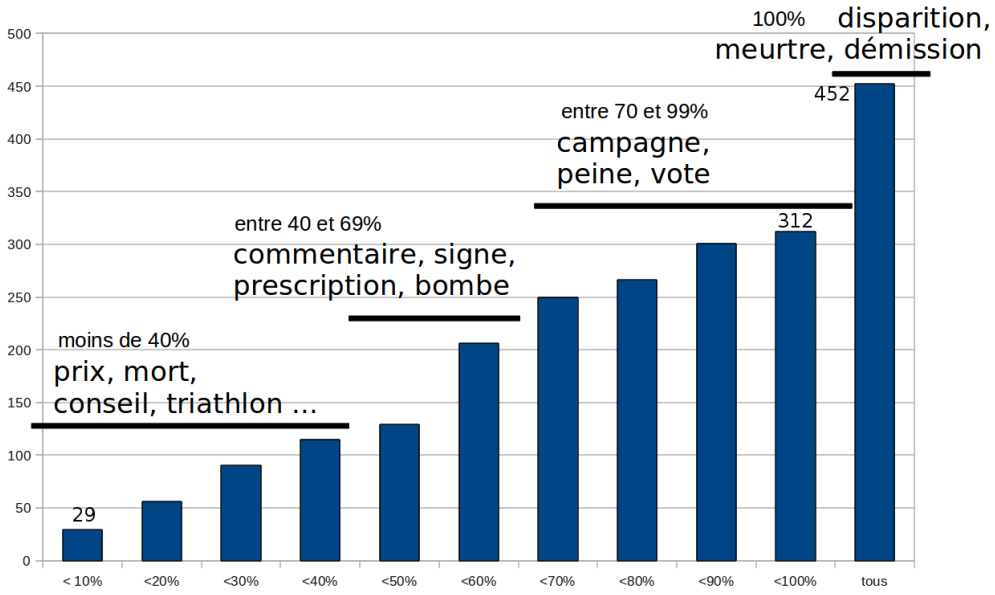

FIGURE 2.2: Progression du nombre de noms qui ont une lecture événementielle par rapport au nombre d'occurrences totales de chaque nom. Par exemple, 29 noms ont une lecture événementielle dans au moins 10 % de leurs occurrences, alors que 312 noms ont une lecture événementielle dans au plus 99 % de leurs occurrences.

(2.14) `Un juge fédéral casse le verdict condamnant Mumia Abu Jamal à la <event> peine de mort </event>.` (verdict)

(2.15) `respecter l'esprit de notre législation française qui n'inclut pas la peine`$_{(abstrait)}$ `de mort "` (sanction)

Au moyen de ces exemples, on se rend compte de manière plus précise de l'ambiguïté causée par la polysémie des mots : "`mort`" désigne l'événement, le processus, mais aussi la personne ou le résultat de cet événement. De plus, certains mots se révèlent dans notre corpus plus événementiels que d'autres : "`campagne`" désignant le lieu ou le paysage rural, aussi bien que l'activité politique ou commerciale ; ce mot est plus souvent événementiel (plus de 70 %) que "`conseil`" qui désigne le groupe de personne ou la séance tenu par ce groupe de personnes (moins de 40 %). Il faut ici remarquer que la nature journalistique du corpus a sans aucun doute une forte influence sur les scores de certains mots, et que des proportions différentes seraient probablement obtenues dans d'autres genres textuels.

2.1.3 Création automatique de lexiques pondérés de noms d'événements

Toutes les utilisations de lexiques en TAL sont sujettes à la problématique de l'équilibre à trouver entre la couverture et l'ambiguïté du lexique. Prenons l'exemple simple d'une liste de prénoms, élément quasiment indispensable pour détecter les entités nommées de type *personne*. Vouloir ajouter tous les prénoms connus, dans toutes les cultures, conduit à une ambiguïté trop importante (entre prénom et nom commun, nom de lieu, nom d'organisation, etc.) et rend finalement la ressource inutilisable. À l'opposé, réduire la liste aux seules entrées non ambiguës réduit trop la couverture du lexique. On en est donc réduit à adopter un niveau "acceptable" de façon souvent empirique.

Le cas des noms d'événements est un cas de figure extrême de ce problème, puisque nous avons vu que quasiment tous les noms ayant une acception événementielle en ont également au moins une autre. C'est pourquoi nous jugeons que les lexiques de noms d'événements existants, notamment le *VerbAction* pour le français [145] et *WordNet* pour l'anglais [114], sont insuffisants. Nous proposons donc une méthode permettant de créer un lexique pondéré des noms d'événements, c'est-à-dire une liste dans laquelle chaque nom est associé à un poids représentant *à quel point* ce mot est susceptible d'avoir une lecture événementielle [13]. La charge de la décision (événement ou pas) devra alors être prise dans le contexte de la phrase, et revient donc au système d'annotation automatique.

Principe

Le principe de construction de ce lexique est relativement simple et peut être appliqué à la génération d'autres types de lexiques pondérés : un exemple de règles de détection de noms d'événements est écrit par un expert humain. Ces règles, potentiellement très peu nombreuses, doivent être conçues de manière à éviter au maximum l'ambiguïté. Par exemple, on peut dire que le sujet du verbe "avoir lieu" est un événement, de façon non ambiguë, et ainsi étiqueter automatiquement toutes les occurrences de ces sujets comme nom d'événement.

De telles règles ont pour objectif une haute précision (peu de faux positifs). En revanche, il est certain que par leur faible nombre, elles conduiront à une couverture très faible (beaucoup de faux négatifs). Pour compenser cette insuffisance et collecter un nombre significatif de mots et d'occurrences de ces mots, il est nécessaire d'analyser un très large volume de documents.

Nous obtenons ainsi l'annotation d'un grand nombre de mots différents. Ces mots sont parfois étiquetés par les règles, mais le plus souvent ne le sont pas. Nous pouvons cependant en déduire un "poids relatif d'événementialité" (ou ERW pour *eventiveness relative weight*) pour chaque nom w, calculé comme le ratio entre le nombre d'occurrences $e(w)$ du nom w extrait par les règles, par le nombre total d'occurrences $t(w)$ de ce mot dans le corpus :

$$ERW(w) = \frac{e(w)}{t(w)}$$

Nous obtenons une valeur d'ERW qui, si elle ne représente bien sûr pas une proportion ou une probabilité de l'usage événementiel du mot, permet, en comparaison des poids de l'ensemble des mots, d'estimer son degré d'ambiguïté relatif. Ceci nous permet en quelque sorte de prédire à quel point un nom peut être événement par rapport à d'autres.

Les règles d'extraction

Toutes nos règles d'extraction ont été codées selon le formalisme de *XIP*, un outil plus général d'analyse linguistique (Xerox Incremental Parser [4]). XIP est un produit du centre de recherche de Xerox (XRCE) qui fournit des analyses syntaxiques profondes, robustes et rapides de textes. Il extrait des relations syntaxiques entre les éléments de la phrase, sous forme de dépendances, mais également des rôles thématiques généraux et des entités nommées. XIP est un système à base de règles et son architecture peut être divisée en trois parties :

– Une étape de prétraitement pour la segmentation des phrases et des mots, puis l'analyse morpho-syntaxique ;

– Une analyse syntaxique de surface avec du *chunking* et un étiquetage des entités nommées ;
– Un traitement plus profond avec une analyse des dépendances syntaxiques génériques (sujet, objet direct, détermination, etc.) puis de structure plus complexes comme les rôles thématiques, les clauses imbriquées, etc.

Nous utilisons l'analyse syntaxique que cet outil nous procure. Ce travail a été fait pour le français et pour l'anglais.

Nous distinguons deux types de règles :

1. Les règles d'extraction fondées sur des critères temporels. En particulier, les noms suivis de prépositions à caractère exclusivement temporel sont étiquetés comme des événements. Ces prépositions indiquent qu'un événement se produit (`à l'occasion de`, `lors de`...) ou réfèrent à un tel événement (`pendant`, `à la suite de`, `à l'issue de`...) :

   (2.16) Les activistes qu'ils ont libérés *au début de* l'*Intifada* [...]

2. Les règles basées sur des indices verbaux introduisant des événements (`se produire`, `avoir lieu`) ou des causes et des conséquences (`entraîner`, `provoquer`, `s'expliquer par`).

   (2.17) `Le` *Sommet du G8* `est organisé à Deauville.`

   Ces règles nécessitent certaines vérifications contextuelles, en particulier sur la nature du nom concerné (pas un nom de personne, d'organisation ou de lieu). Même si elles sont plus ambiguës que les précédentes, nous avons sélectionné par une étude de corpus [11] les verbes qui introduisaient des événements dans plus de 90 % des cas, ce qui permet de conserver une très bonne précision.

Les lexiques

Nous avons utilisé les règles décrites ci-dessus sur des corpus de l'Agence France Presse (AFP), en français (390 millions de mots) et en anglais (427 millions de mots).

Les tableaux 2.1 et 2.2 proposent des examples de mots présents dans les lexiques pondérés. Les tableaux sont doubles, la partie a. englobant les lemmes qui sont présents dans les lexiques standards (*VerbAction* pour le français et les *actions* et les *événements* de *WordNet* pour l'anglais) et la partie b. pour les lemmes étrangers à ces lexiques.

On voit que les noms peu ou pas du tout ambigus (ceux qui sont toujours des amorces de noms d'événements) ont un ERW relativement élevé. C'est le cas de "`chute`", "`élection`", "`krach`", "`overthrow`", "`breastfeeding`". On note que "`clôture`", fortement ambigu dans le cas général, l'est semble-t-il beaucoup moins dans ce type de corpus (la presse), où l'objet est bien moins évoqué que le fait de clore. En revanche, des mots comme "`tension`", "`subvention`" ou "`accès`" sont très ambigus et obtiennent une valeur d'ERW peu élevée. C'est également le cas de la date "`11 septembre`", mais celle-ci est une des seules dates du lexique obtenu, et a de loin le meilleur ERW. En effet, même si cette date est souvent utilisée, le contexte événementiel est activé plus souvent que les autres dates.

2.1.4 Du verbe au nom : évolution des désignations d'événements

Pour conclure cette série d'études sur la désignation nominale des événements, intéressons-nous maintenant à l'évolution des désignations d'événements [147] [1]. En ef-

1. Travail effectué avec Béatrice Arnulphy, Ruixin He et Véronique Moriceau.

	Nb. détecté par les règles	Nb. total d'occurrences	**ERW**
chute	434	2620	**0,166**
clôture	63	470	**0,134**
élection	1243	9713	**0,128**
guerre	1126	11542	**0,098**
crise	286	6185	**0,046**
expérience	63	2878	**0,022**
tension	16	1595	**0,001**
coopération	5	1631	**0,003**
subvention	2	867	**0,002**

a. Mots appartenant aux lexiques standards

	Nb. détecté par les règles	Nb. total d'occurrences	*ERW*
Anschluss	3	4	**0,750**
méchoui	3	5	**0,600**
krach	20	169	**0,118**
RTT	14	166	**0,084**
demi-finale	35	553	**0,063**
cessez-le-feu	15	440	**0,034**
difficulté	16	3894	**0,004**
accès	9	2828	**0,003**
11 septembre	12	4354	**0,003**

b. Mots absents des lexiques standards

TABLE 2.1: Exemples de mots collectés par les règles d'extraction en français.

fet, la façon de désigner un même événement évolue avec le temps dans les textes, en particulier dans les médias. Les premiers récits utilisent généralement une description verbale (`Des avions se sont écrasés sur les tours du World Trade Center à New York`), avant de nominaliser cette description (`l'attentat de mardi à New York`) puis de la raccourcir (dans ce cas extrême, par une simple date `11-Septembre`, c'est ce qu'on appelle un héméronyme [30]).

Cette problématique est bien sûr importante en extraction et en recherche d'information, où conserver un "fil" des références à un même événement pourrait augmenter les performances des systèmes.

Nous avons donc souhaité valider par une étude de corpus un certain nombre d'hypothèses concernant l'évolution des désignations d'événements [147]. Pour ce faire, nous avons composé un corpus composé de sept ensembles distincts de phrases, chacun contenant des références à un événement défini à l'avance (soit sept événements de granularités variables : *la crise grecque*, *la révolution tunisienne*, *la révolution égyptienne*, *l'affaire Wikileaks*, *l'affaire Laëtitia*, *la grippe H1N1*, *le nuage de cendres islandais*).

Pour chacun de ses sujets, le sous-corpus a été composé en collectant des documents sur chaque thème par des requêtes à un moteur de recherche (*Lucene*), puis en sélectionnant manuellement les passages de ces documents contenant des désignations des événements concernés. Chaque phrase a ensuite été annotée avec l'outil Callisto [45] pour distinguer les désignations verbales des désignations nominales. Ceci permet de

	Traduction française	Nb. détecté par les règles	Nb. total d'occurrences	**ERW**
overthrow	renversement	383	448	**0,855**
intifada	Intifada	7	11	**0,636**
bombardement	bombardement	6	12	**0,500**
testimony	testament	426	13109	**0,032**
sleepover	soirée pyjama	3	27	**0,111**
publication	publication	154	9337	**0,016**
marathon	marathon	52	8070	**0,006**

a. Mots appartenant au lexique existant

	traduction française	Nb. détecté par les règles	Nb. total d'occurrences	**ERW**
play-off	barrages	73	75	**0,973**
breastfeeding	allaitement	3	4	**0,750**
overheat	surchauffe	3	7	**0,428**
stopover	arrêt	372	1345	**0,276**
cross-examination	examen croisé	53	416	**0,127**
distillery	distillerie	4	126	**0,032**
welcome	bienvenue	66	3884	**0,017**
influenza	grippe	37	6019	**0,006**

b. Mots absents des lexiques standards

TABLE 2.2: Exemples de mots collectés par les règles d'extraction en anglais.

comparer la nature des désignations, ainsi que leur longueur et leur variété, dans le temps.

Cette première étude a conduit aux conclusions suivantes :

1. Après de premières descriptions verbales juste après l'événement, celui-ci prend part au contexte d'événements ultérieurs, et est alors nominalisé, sauf si un laps de temps important s'écoule sans en faire mention. L'exemple de la révolution tunisienne (figure 2.3) l'illustre bien, en montrant l'évolution dans le temps du nombre de désignations verbales (courbe rouge et continue) ou nominales (courbe à tirets bleus).

   Le point ① dans la figure correspond à la fuite du président Ben Ali (14 janvier 2011), précédée par une progression de désignations illustrant l'aggravation de la situation. Les jours suivants, les désignations nominales prennent le relai.

2. Les journalistes appliquent un principe d'économie conduisant, au fil des jours, à des désignations plus courtes et à des variations moins nombreuses. Ces deux phénomènes peuvent être illustrés par la polémique politico-judiciaire sur le meurtre de Laëtita Perrais en 2011, qui a continué jusqu'en 2013. On a d'abord parlé de "`la disparition d'une jeune femme dans la nuit de mardi à mercredi dans la ville de Pornic (Loire-Atlantique)`", puis de "`la disparition de Laëtitia à Pornic`", "`la disparition de Laëtitia`" avant que la presse ne converge 11 jours après les faits vers "`l'affaire Laëtita`".

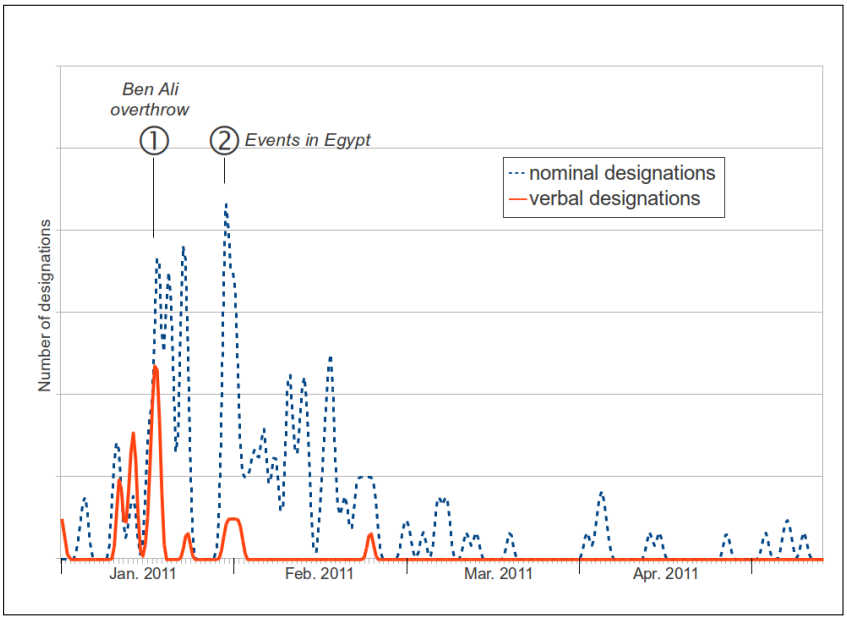

FIGURE 2.3: Évolution des désignations d'événements, l'exemple des événements en Tunisie en 2011.

2.1.5 Conclusion

Outre la perspective d'un meilleur étiquetage des événements nominaux dans les textes, confirmée très récemment par [9], ces études sur les désignations nominales des événements sont intéressantes à plusieurs titres. Tout d'abord, le domaine général de l'extraction d'information peut en bénéficier. La notion d'événement est souvent centrale dans ce domaine, et nommer les événements de façon plus précise peut améliorer le regroupement des informations et sa compréhension par l'utilisateur final. De plus, elles ouvrent la voie vers une meilleure résolution des co-références d'événements, problème complexe demandant de nombreuses connaissances [21, 97, 42]. Enfin, elles peuvent contribuer à simplifier ou élargir les études linguistiques ou journalistiques sur des événements précis, jusqu'ici réalisées en étudiant les textes de façon totalement manuelle [59, 30, 96, 31].

2.2 Identification des relations temporelles entre événements

Nous allons aborder la question de l'identification des relations temporelles par la présentation de deux approches radicalement opposées et mises en œuvre à des périodes très différentes. La première, présentée dans cette section, concerne des documents journalistiques du domaine général, et utilise des méthodes symboliques à bases de règles, à travers l'outil d'analyse XIP. La seconde, décrite à la section suivante, est une application à des textes médicaux, entièrement basée sur des méthodes d'apprentissage automatique.

La tâche de détection des relations temporelles dans les textes est complexe et reste d'ailleurs un défi, pour plusieurs raisons :

– L'information temporelle est relayée dans les textes par de nombreuses sources différentes (connaissance de la sémantique lexicale, aspects syntaxiques, modes et temps grammaticaux...) ;

– Des connaissances extra-linguistiques sont nécessaires pour ordonner les événements correctement (par exemple, dans "`il a ouvert la porte et est sorti`", la connaissance du monde nous dit que le premier événement a nécessairement précédé l'autre, tandis que dans "`Il a mangé et bu`", aucun ordre temporel n'est impliqué) ;
– Du raisonnement est nécessaire pour déduire certaines relations, notamment parce que celles-ci sont transitives.

Telle que formalisée par la campagne TempEval [156] et décrite au chapitre 1, cette tâche est composée de trois parties :

– A : identification des relations temporelles entre les événements d'une même phrase ;
– B : identification de la relation temporelle entre chaque événement et la date de création du document ;
– C : identification des relations temporelles entre événements de phrases adjacentes.

Les figures 2.4 et 2.5 rappellent le type de résultat attendu, au format TimeML [128] et avec une représentation graphique.

Nous présentons donc dans cette section la première approche, dans laquelle l'analyse temporelle est intégrée dans l'outil XIP, décrit ci-dessus.

Le traitement temporel des textes est intégré dans les règles de XIP, au niveau de la phrase, au même niveau que les autres traitements linguistiques. L'association entre des expressions temporelles et des événements, base de l'analyse, est considérée comme un cas particulier de la tâche plus générale d'attachement des rôles aux prédicats. En retour, nous avons montré qu'un traitement adéquat des expressions temporelles bénéficie à la tâche d'analyse syntaxique, car elle permet parfois d'éviter des erreurs de regroupement en syntagmes [1].

2.2.1 Événements

Nous avons déjà eu l'occasion de montrer qu'il était difficile de décider de ce qu'était un événement. Dans le cas de l'identification des relations temporelles, la communauté s'est clairement orientée vers la définition du standard TimeML, et voit l'événement comme le mot dans la phrase qui porte la description de l'événement. La question des noms d'événement se pose donc tout particulièrement ; les travaux décrits dans cette section étant antérieurs à ceux sur la classification automatique des noms d'événement, nous avons adopté l'approche suivante, plus heuristique. Est considéré comme un événement :

– Tous les verbes, y compris les verbes d'état (suivant la définition de TimeML) ;
– Tous les noms déverbaux, ou en tout cas morphologiquement lié à un verbe (XIP fournissant cette information) ;
– Tout nom appartenant à une liste prédéfinie, composée pour l'occasion (`guerre`, `festival`, etc.). Cette liste a été obtenue en analysant un large corpus de l'agence Reuters et en sélectionnant les noms précédés de la préposition temporelle *during* (la seule préposition exclusivement temporelle en anglais), ou sujets des verbes *to last*, *to happen*, *to occur*.

1. Travail réalisé avec Caroline Hagège.

```
<TEXT>
    <TIMEX3 functionInDocument="CREATION_TIME" tid="t1"
    type="DATE" value="1999-05-19"/>

    Le président russe Boris Eltsine a <EVENT class="I_ACTION" eiid="ei1"
    pos="VERB" pred="ANNULER" tense="PAST" vform="PASTPART">annulé</EVENT>
    <TIMEX3 temporalFunction="true" tid="t2" type="DATE" value="1999-05-18"
    valueFromFunction="tf17">hier</TIMEX3> une <EVENT class="OCCURRENCE"
    eiid="ei2" pos="NOUN" pred="RENCONTRE">rencontre</EVENT>
    avec le président du gouvernement espagnol Jose
    Maria Aznar en raison d'une <EVENT class="STATE"
    eiid="ei14" pos="NOUN" pred="BRONCHITE">bronchite</EVENT>, a
    <EVENT class="REPORTING" eiid="ei4" pos="VERB" pred="AFFIRMER" tense="PAST"
    vform="PASTPART">affirmé</EVENT> l'ambassade d'Espagne à Moscou.
</TEXT>
<TLINK lid="l1" relType="BEFORE" relatedToTime="t1" timeID="t2"/>
<TLINK eventInstanceID="ei1" lid="l4" relType="IS_INCLUDED" relatedToTime="t2"/>
<TLINK eventInstanceID="ei4" lid="l6" relType="BEFORE" relatedToTime="t1"/>
<TLINK eventInstanceID="ei14" lid="l10" relType="INCLUDES" relatedToTime="t2"/>
```

FIGURE 2.4: Exemple d'annotation des événements et des expressions temporelles, au format TimeML (XML). Une représentation graphique plus lisible est présentée à la figure 1.1.

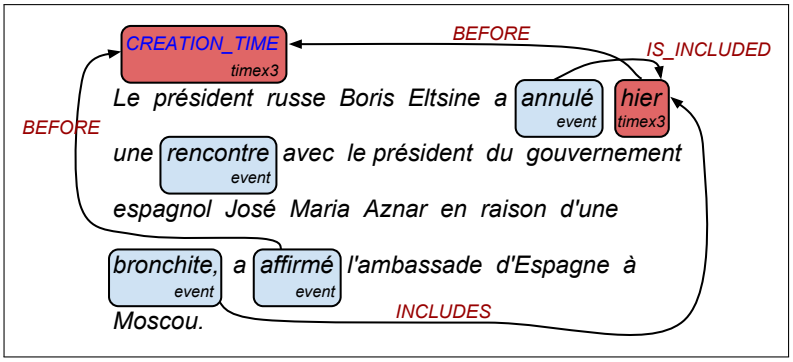

FIGURE 2.5: Exemple d'annotation des événements et des expressions temporelles, représentation graphique. La version textuelle annotée en XML est présentée à la figure 2.4.

2.2.2 Analyse temporelle

Nous pouvons distinguer dans notre système trois niveaux principaux d'analyse des expressions temporelles de détection des relations. Au niveau local, on reconnaît et on analyse les expressions temporelles. Au niveau de la phrase, on les attache aux événements qu'elles modifient, et on propose un ordre pour les événements apparaissant dans la même phrase. Enfin, au niveau du document, on ordonne les événements lorsque cela est possible.

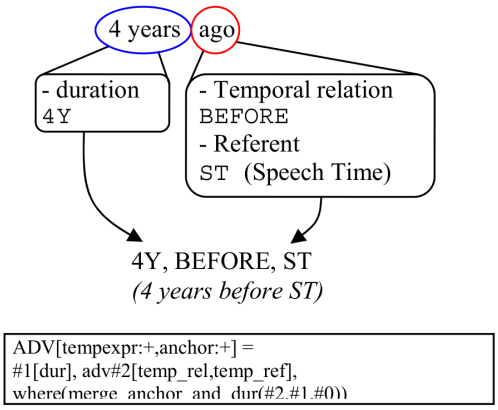

FIGURE 2.6: Analyse d'une date relative au niveau local (ST = Speech Time).

Analyse locale

Si la reconnaissance des expressions temporelles est un problème partiellement résolu (avec des taux de reconnaissance tournant autour des 85 % en anglais), certains problèmes de segmentation se posent pour les expressions complexes comme "`deux jours après Noël`", "`pendant 10 jours, en septembre`", etc. La question se pose de savoir si de telles expressions doivent être considérées comme une ou comme plusieurs expressions. Les deux critères, l'un syntaxique, l'autre sémantique, que nous avons alors appliqués, sont les suivants. Une expression temporelle doit être séparée en deux expressions minimales si :

1. La phrase reste syntaxiquement valide si une seule des deux expressions est conservée ;
2. La combinaison événement-expression minimale peut être inférée de la combinaison événement-expression complexe.

Ainsi, dans la phrase

(2.18) Nous nous rencontrons (deux fois) (chaque année).

L'expression "`deux fois chaque année`" est une expression unique, car la condition 2 n'est pas satisfaite ("`Nous nous rencontrons deux fois`" ne correspond pas à la réalité introduite par la phrase complète). En revanche, dans

(2.19) Nous avons voyagé (pendant 10 jours) (au mois de septembre).

Il est nécessaire de séparer les deux expressions, car les deux conditions sont vérifiées ("`Nous avons voyagé pendant 10 jours`" et "`Nous avons voyagé au mois de septembre`" sont sémantiquement compatibles avec la phrase complète).

Nous avons une approche compositionnelle de la sémantique des expressions temporelles, en reconnaissant les expressions temporelles par des règles contextuelles et en attribuant des actions à chaque regroupement. Ceci implique de considérer les signaux (prépositions ou adverbes temporels) comme partie intégrante de l'expression, en opposition au standard TimeML [128].

La figure 2.6 illustre cette étape avec un exemple pour une date relative très simple. La règle construit un nœud ADV (adverbial) avec des traits booléens associés (à gauche du signe '=') à partir de l'expression "`4 years ago`" (à droite). L'action associée est une

fonction nommée `merge_anchor_and_dur` dont les paramètres sont des nœuds syntaxiques.

Au niveau de la phrase

Au-delà du syntagme, des relations entre les éléments sont établies, représentées par des dépendances syntaxiques ou sémantiques (par exemple, la dépendance *SUBJECT* entre un groupe verbal et un groupe nominal). C'est à ce niveau que les liens entre prédicats et expressions temporelles sont définis, ainsi que les relations temporelles entre événements d'une même phrase (tâche A), souvent dans des propositions imbriquées. Les exemples suivants illustrent le type de relations extraites :

(2.20) Les habitants ont commencé à rejoindre Abuja mardi pour un rassemblement de deux jours.

⇒ `DATE(mardi)`, `DURATION(deux jours)`
`EVENT(commencé)`, `EVENT(rejoindre)`, `EVENT(ressemblement)`
`ORDER[starts](commencé, rejoindre)`
`TEMP(commencé, mardi)`
`TEMP(rassemblement, deux jours)`

(2.21) Après 10 ans d'expansion, ils parlent maintenant de licenciements.

⇒ `DURATION(10 ans)`
`EVENT(expansion)`, `EVENT(licenciements)`
`ORDER[before](expansion, licenciements)`
`TEMP(expansion, 10 ans)`

(2.22) Ce changement vient un mois après la suspension de plusieurs lignes par Qantas.

⇒ `DURATION(un mois)`
`EVENT(vient)`, `EVENT(suspension)`
`ORDER[before](suspension, vient)`
`DELTA(suspension, vient, un mois)`

On voit donc comment on obtient à ce niveau des premières relations temporelles entre des événements portés par des propositions subordonnées ou coordonnées de la même phrase (dépendance `ORDER`).

Dans le dernier exemple, la dépendance `DELTA` exprime en plus que l'intervalle entre les deux événements est d'un mois, ce qui permettra éventuellement de découvrir d'autres relations si des événements dans cette période sont détectés.

À ce niveau, des informations lexicales (le verbe `commencer` introduit la relation `starts`) sont utilisées, ainsi que des informations morphologiques sur les temps grammaticaux, notamment pour obtenir la relation par rapport au moment de la création du document (tâche B). Ainsi, un verbe au passé désigne un événement ayant eu lieu avant la création du document.

Au niveau du document

Trouver les relations temporelles au niveau du document est un problème non résolu, y compris pour les phrases adjacentes (tâche C). Certaines connaissances permettent néanmoins de proposer un ordonnancement partiel. En particulier, grâce à la connaissance de la date de création du document, il est possible de normaliser certaines

expressions temporelles relatives (comme "`il y a deux jours`" ou "`mardi dernier`"), et en les reliant à des événements, à ordonner ceux-ci. De plus, les temps des verbes permettent de fournir certaines informations supplémentaires (entre passé et futur, bien sûr, mais également par exemple entre passé simple et plus-que-parfait [120]). Enfin, la transitivité des relations temporelles (par exemple, *A avant B et B avant C* $\Rightarrow$ *A avant C*) nous permet d'inférer un maximum de relations temporelles avec des règles de composition, dans le but de saturer le graphe des relations [6].

2.2.3 Discussion

Les approches symboliques pour la détection et la normalisation des expressions temporelles restent parmi les plus performantes à l'heure actuelle [152]. En revanche, l'induction des relations temporelles entre les événements a, depuis ces travaux réalisés en 2004 puis 2007, progressé de façon exclusive à l'aide de techniques d'apprentissage supervisé, à l'exception des relations vraiment lointaines (deux événements distants dans le texte) dont l'extraction peut tirer partie de règles simples et pour lesquelles les classifieurs automatiques restent très peu performants. Ces approches ont notamment bénéficié des efforts constants d'annotation semi-automatique de corpus, en anglais et dans plusieurs autres langues, dont le français [152].

C'est en s'inscrivant dans cette mouvance que nous avons abordé la question spécifique des relations temporelles dans les textes du domaine médical [66, 168] [1].

2.3 Application au domaine médical

Le traitement automatique des langues dans le domaine bio-médical s'est également penché sur la question des événements et de leurs relations, notamment au travers de la campagne d'évaluation i2b2 en 2012 [144], dans le but de créer des outils d'analyse des comptes-rendus d'hospitalisation, détaillant l'évolution d'un patient et de ses problèmes.

Outre les spécifités classiques des travaux sur les domaines techniques (vocabulaire, sémantique, recours à des ressources terminologiques, etc.), deux aspects importants, déjà mentionnés au chapitre 1, conduisent à des approches différentes de celles employées dans le domaine général :

- Les premiers travaux sur le sujet ont conduit à éloigner l'événement de la notion de prédicat pour le rapprocher des *concepts* médicaux. En conséquence, les verbes et les noms prédicatifs ne sont plus les annonceurs privilégiés des événéments, au profit de noms de problèmes médicaux, de traitements (médicaments), d'examens ou même de lieux.
- La structuration quasi-systématique, et souvent explicite, des comptes-rendus autour de deux dates (hospitalisation – *Admission Date*, ou AD – et sortie – Discharge Date, ou DD) et de grandes phases du parcours du patient (histoire de la maladie – HPI pour *History of Present Illness*, évolution dans le service – HC pour *Hospital Course*), peut guider l'induction des relations temporelles.

La tâche en elle-même, telle que mise en œuvre pour la campagne d'évaluation i2b2, diffère de la tâche TempEval présentée plus haut de plusieurs façons. En particulier, elle ne restreint pas les situations où l'on peut trouver ces relations : dépendance syntaxique ou pas, phrases consécutives ou pas, etc. Pour rappel, un compte rendu typique est présenté partiellement à la figure 2.7, avec les annotations sur les événements et les expressions temporelles, et finalement les relations à trouver.

1. Travail réalisé avec Pierre Zweigenbaum.

ADMISSION DATE : <TIMEX3 type="date" id="t1">**10/17/95**</TIMEX3>
DISCHARGE DATE : <TIMEX3 type="date" id="t2">**10/20/95**</TIMEX3>
HISTORY OF PRESENT ILLNESS :
This is a 73-year-old man with <EVENT type="problem" id="e1">**squamous cell carcinoma of the lung**</EVENT>, status post <EVENT type="treatment" id="e2">**lobectomy**</EVENT> and <EVENT type="treatment" id="e3">**resection**</EVENT> of <EVENT type="problem" id="e4">**left cervical recurrence**</EVENT>, <EVENT type="occurrence" id="e5">**admitted**</EVENT> here with <EVENT type="problem" id="e6">**fever**</EVENT> and <EVENT type="problem" id="e7">**neutropenia**</EVENT>.
(...)

HOSPITAL COURSE :
He was started on <EVENT type="treatment" id="e8">**Neupogen**</EVENT>, 400 mcg. subq. q.d.
(...)
He was <EVENT type="occurrence" id="e9">**discharged**</EVENT> home on <EVENT type="treatment" id="e10">**Neupogen**</EVENT>.

```
e3 OVERLAP e4      e1 OVERLAP t1      e2 BEFORE t1      e3 BEFORE t1
e4 BEFORE t1       e6 OVERLAP t1      e7 OVERLAP t1     e8 BEFORE t2
e9 OVERLAP t2      e10 OVERLAP t2
```

FIGURE 2.7: Extrait d'article (anonymisé) du corpus i2b2/VA, montrant les événements (EVENT), les expressions temporelles (TIMEX3) ainsi que quelques relations temporelles.

2.3.1 Méthode

Pour identifier les relations temporelles entre les événements (EVENT) et les expressions temporelles (TIMEX), nous avons utilisé la structure des documents en segmentant l'espace de détection en 57 situations distinctes, combinaisons des dimensions suivantes appliquées sur des paires d'éléments (nous utilisons le terme EVT pour désigner indifféremment un EVENT ou TIMEX) :

- Section des EVT source et cible : *AD, DD, HPI, HC* ;
- Types d'éléments des EVT source et cible : *TIMEX* ou *EVENT* ;
- Distance entre EVT : *même phrase (SS), phrases adjacentes de la même section (S1), phrases distantes* ;
- Nombre de TIMEX entre deux événements : aucune (NTB) ou au moins une.

Ainsi, *TIMEX-EVENT-DD-HC* formalise une situation avec une expression temporelle (TIMEX) dans la section « Discharge Date » (DD) et un événement (EVENT) dans la section « Hospital Course » (Figure 2.8). Ces distinctions sont importantes dans la mesure où, par exemple dans cette situation, la plupart des événements cités dans le séjour (HC) sont antérieurs à la date de sortie. Pour chaque combinaison, une méthode appropriée sera donc appliquée.

Pour toutes les paires du corpus d'apprentissage correspondant à une situation, nous avons testé plusieurs classifieurs pour assigner à chaque paire l'une des quatre classes. Cette approche peut être comparée à un arbre de décision initial dont les caractéristiques seraient fondées sur ces quatre dimensions et pour lequel un autre classifieur serait appliqué à chaque feuille. Ainsi, une décision par défaut pour *TIMEX-EVENT-DD-HC* pourrait être que l'événement intervient avant la date de sortie.

La question qui se pose alors est d'identifier les situations les plus importantes à traiter et les méthodes à utiliser. Nous avons pour cela calculé la proportion de

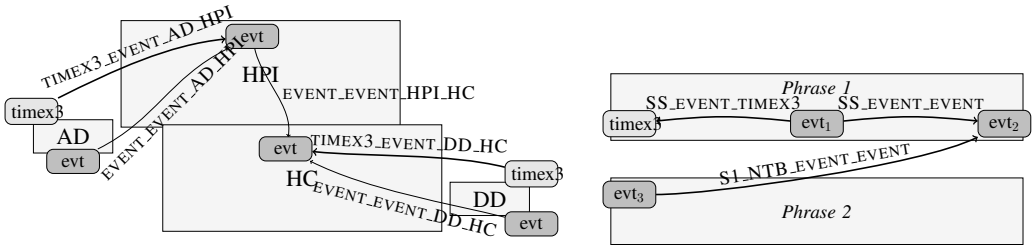

FIGURE 2.8: Exemples de situations, liées à la période de la section (gauche : *_AD_HPI, *_DD_HC), dans la phrase courante ou dans la suivante (droite, SS*, S1*)

relations dans l'ensemble du corpus de référence correspondant à chaque situation. Ceci constitue le rappel idéal qui devrait être atteint par un classifieur entraîné pour cette situation (figure 2.9). Nous nous sommes focalisés sur les situations qui représentent plus de 1 % des cas. Pour ces principales situations, nous avons évalué plusieurs types de classifieurs par validation croisée (arbres de décision J48 de Weka, bayésien naïf, SVM, k plus proches voisins, MaxEnt, forêt d'arbres décisionnels, régression logistique). Nous avons ensuite retenu le meilleur classifieur pour chaque situation, et expérimenté également des procédures de vote entre classifieurs.

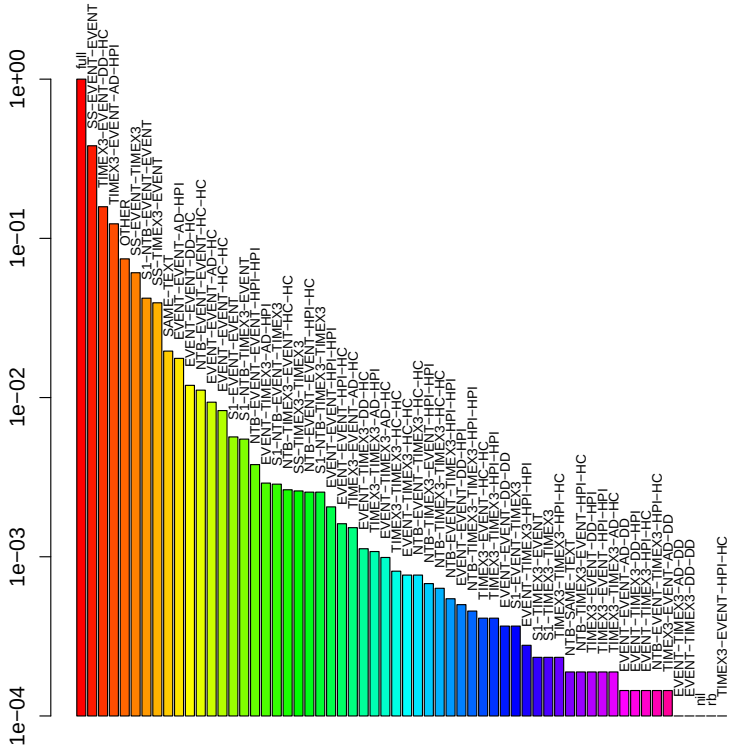

FIGURE 2.9: Répartition du nombre de relations pour chaque situation dans le corpus d'apprentissage (échelle semi-logarithmique), indiquant le rappel maximum atteignable pour chaque classifieur correspondant.

2.3.2 Caractéristiques d'apprentissage, résultats et discussion.

Les caractéristiques choisies pour l'apprentissage sont à la fois internes et contextuelles :

- Caractéristiques internes :
  - Toutes les annotations de l'EVT (modalité, polarité, sous-types) ;
  - Pour les événements, une sous-catégorisation d'après une étude du lexique dans le jeu d'apprentissage *(chirurgie, état, événement ponctuel, localisation, suivi d'événement, autres)* ;
  - Le texte de l'EVT (si parmi les 50 les plus fréquents) ;
  - Les mots de l'EVT (si présents au moins 10 fois dans le corpus d'apprentissage) ;
  - Nous avons également testé des classes distributionnelles (clusters de Brown) sur les mots des EVT, combinées aux prépositions temporelles, mais cela n'a permis aucune amélioration.
- Caractéristiques contextuelles :
  - La distance entre les deux EVT de la relation ;
  - La présence de prépositions temporelles ou non entre EVT ;
  - Le nombre d'autres EVT entre les deux ;
  - Pour les paires d'événements dans une même phrase, les dépendances syntaxiques obtenues par l'analyseur syntaxique Charniak McClosky converties en dépendances de Stanford ;
  - Des combinaisons de ces traits, par exemple *<type-EVT-source, dépendance-syntaxique, type-EVT-cible>*, où les types associés à *source* et *cible* sont les catégories i2b2 *(clinical department, evidential, occurrence, problem, test, treatment ; date, duration, frequency, time)*.

D'autre part, certaines relations peuvent être identifiées à partir de simples règles : l'admission est avant la sortie, la date et les événements liés à l'admission se chevauchent.

Nous nous sommes intéressés aux huit situations les plus fréquentes (figure 2.9), abandonnant ainsi d'emblée un rappel de 17 %. La figure 2.10 donne les exemples des résultats obtenus pour les classifieurs J48 (gauche) et par régression logistique (droite).

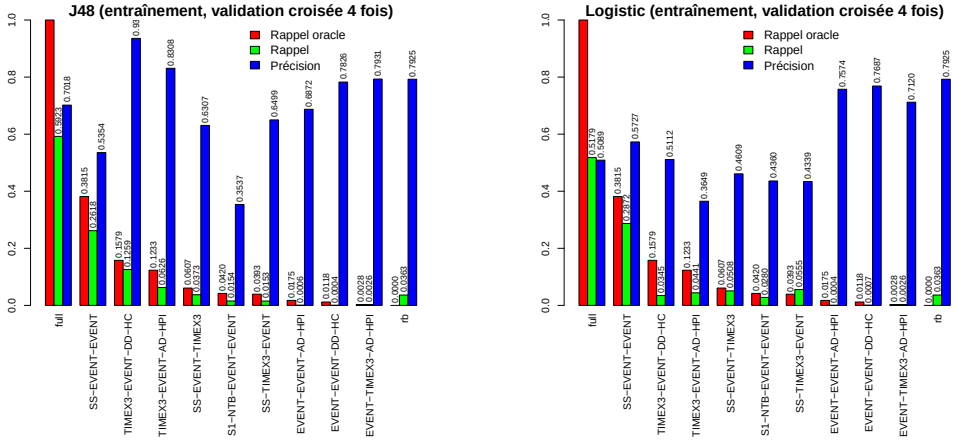

FIGURE 2.10: Performance de J48 sur les 9 meilleures situations (centre), performance de la régression logistique sur les 9 meilleures situations (droite). Les mesures d'évaluation utilisées sont décrites par [144].

Concernant les situations moins fréquentes, nous ne sommes pas parvenus à augmenter le rappel sans de lourdes pertes en précision.

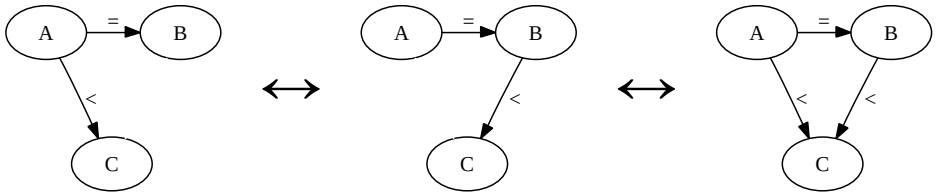

FIGURE 2.11: Trois annotations identiques. La dernière est le résultat de la fermeture transitive.

Dans ce travail, très différent par les approches de celui mené bien plus tôt sur les relations temporelles dans le domaine général (section 2.2), nous avons montré comment une étude systématique des situations où l'on peut rencontrer des relations temporelles dans des comptes rendus hospitaliers nous a permis d'obtenir des résultats sensiblement meilleurs que ceux obtenus sans effectuer cette étude. Cependant, prendre chaque situation de manière indépendante oblige à résoudre des conflits après coup (les relations temporelles étant transitives, des choix sur une situation ont une influence sur d'autres situations), dans notre cas à l'aide de règles de priorité basée sur la meilleure précision obtenue par les classifieurs par validation croisée. Une procédure de décision plus globale, basée sur une étude du graphe produit et sur la confiance des prédiction, puis une sélection du sous-graphe global optimal, pourrait permettre de surmonter cet inconvénient et d'améliorer les performances.

2.4 Métriques d'évaluation

Le dernier point que nous abordons au sujet de l'extraction des événements et des relations temporelles dans les textes est celui de l'évaluation [1]. Les relations temporelles étant transitives, il existe plusieurs façons de représenter un même ordonnancement d'événements (comme le montrent de façon très simplifiée les trois graphes de la figure 2.12). Dans ces conditions, il est impossible de segmenter l'information temporelle en sous-parties indépendantes, en considérant chaque relation temporelle comme une unité d'information que l'on évaluerait séparément des autres, avant de faire une somme et de présenter une mesure classique de précision et de rappel. Dans un graphe temporel donné, étiqueté par des relations d'Allen (figure 1.3 page 8) ou équivalentes, il est indispensable d'opérer dans un premier temps une *fermeture transitive*, c'est-à-dire de compléter le graphe avec toutes les relations induites par celles présentes initialement (dans la figure 2.12, il s'agirait du graphe de droite).

Pourtant, le problème n'est pas résolu pour autant, puisque les informations présentes dans le graphe sont alors très redondantes. Considérons les graphes exemples de la figure 2.12, dans lequel le graphe K est la référence. S_1 contient deux relations, contre six dans K. Il serait pourtant injuste de lui attribuer un score de rappel de $\frac{2}{6}$, puisqu'en pratique, une seule relation est manquante ($B < C$), les autres étant déduites par transitivité. On souhaiterait plutôt un rappel d'environ $\frac{2}{3}$.

Cependant, même une distinction entre les relations "majeures" (en traits pleins dans la figure) et les relations "mineures" (c'est-à-dire inférables automatiquement, en pointillés), ne serait pas suffisante. Ainsi, le graphe S_2 contient la relation $B < D$, qui est mineure dans la référence K, puisque déduite des autres, mais pas dans S_2. Elle apporte donc de l'information et doit être récompensée. Pourtant, on souhaiterait

1. Travail réalisé avec Philippe Muller.

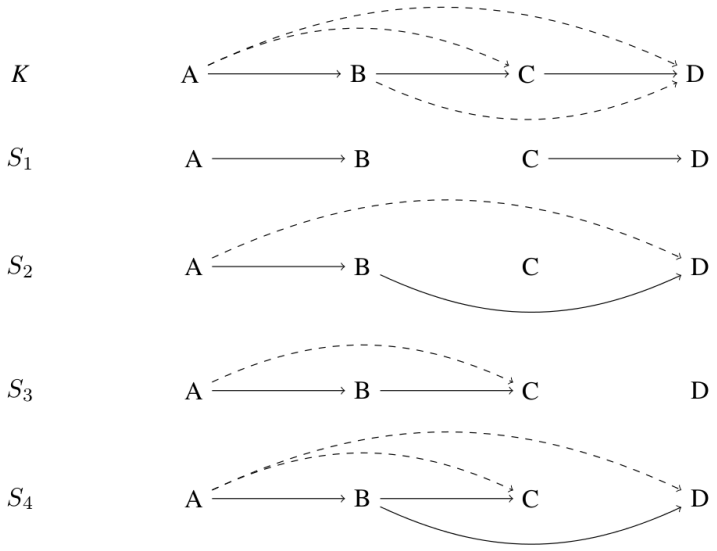

FIGURE 2.12: L'objectif d'une métrique d'évaluation est de considérer équitablement ces différentes configurations.

intuitivement que S_2 obtienne un score plus bas que S_3, malgré une configuration qui semble identique, puisqu'il manque deux relations à S_2 pour parvenir à la référence ($B < C$ et $C < D$), alors qu'il n'en manque qu'une à S_3 ($C < D$).

On voudrait donc une métrique d'évaluation des graphes temporels qui produisent des scores évoluant de façon proportionnelle à l'information temporelle qu'ils contiennent par rapport à la référence, ce qui n'est pas le cas des métriques qui existaient auparavant, qui échouaient à traiter les problèmes de redondance de l'information.

2.4.1 Métrique proposée

L'approche proposée [148] est alors basée sur deux idées initiales :

1. Travailler sur des graphes de points et non plus des graphes d'intervales. Au lieu de manipuler des nœuds représentant des événements composés d'un début et d'une fin, on traite deux nœuds par événement (début et fin). On regroupe également les nœuds représentant des points simultanés. On obtient des graphes tels que celui représenté à la figure 2.13.

   Ceci permet de n'avoir plus que des relations d'égalité et de précédence dans le graphe (au lieu des 13 relations d'Allen auparavant), mais surtout cela permet de produire un graphe minimal unique.

   Un graphe minimal est un graphe dans lequel aucune relation ne peut être enlevée sans perdre de l'information (autrement dit, ce graphe ne contient plus aucune information redondante). La réduction d'un graphe d'événements en un graphe minimal ne produit pas de solution unique [131], tandis que nous avons démontré l'unicité du graphe minimal pour le graphe de points. Il devient ainsi possible de comparer deux graphes en considérant l'ensemble des informations temporelles qu'ils contiennent tout en évitant les problèmes liés à la redondance des informations.
2. Comme pour les mesures d'édition, on évalue un graphe hypothèse en estimant le nombre d'opérations à effectuer pour parvenir au graphe référence. Dans notre cas,

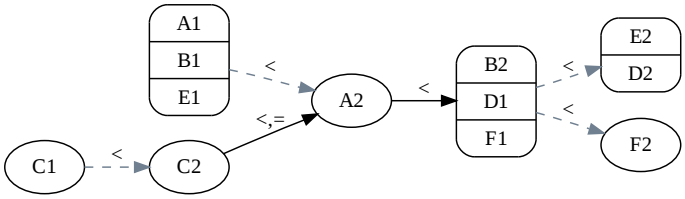

FIGURE 2.13: Exemple de graphe temporel à base de points. Un événement E est constitué d'un nœud de début (E1) et d'un nœud de fin (E2). Ici, A1, B1 et E1 sont simultanés, les nœuds sont donc regroupés en un seul. Les seules relations existant donc dans le graphe sont des relations de précédence, ainsi que, dans les graphes de relations non entièrement spécifiées, des relations "avant ou simultané".

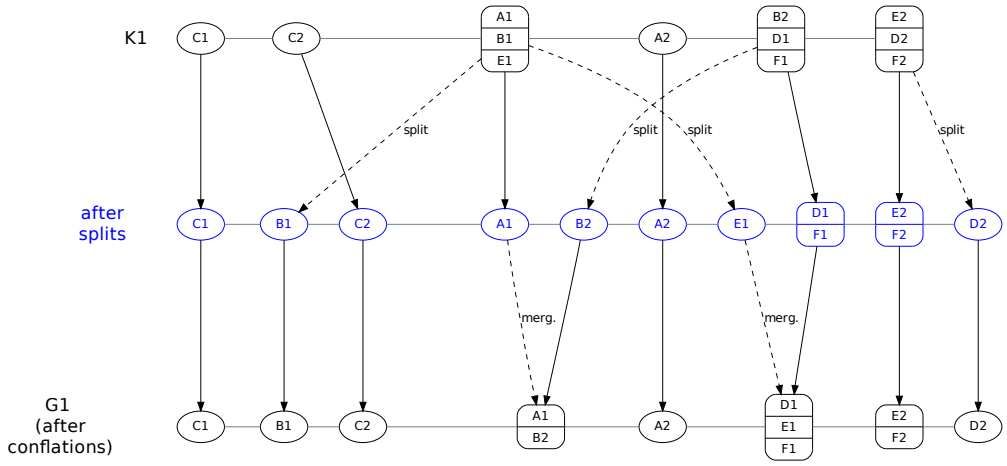

FIGURE 2.14: Opérations de fission (*split*) et de fusion (*conflation*) conduisant d'un graphe K1 à un graphe G1.

les opérations possibles sont la fission d'un nœud (quand deux points simultanés dans un graphe deviennent distincts dans l'autre) ou la fusion d'un nœud (quand deux points liés par la relation de précédence dans un graphe sont regroupés dans l'autre). La figure 2.14 donne un exemple de telles opérations.

2.4.2 Stratégie d'évaluation de la métrique

Il est intéressant de confronter cette nouvelle métrique à des données pour constater que les problèmes cités plus haut ne l'affectent pas. En particulier, nous faisons varier la quantité d'information présente dans les graphes pour montrer que la métrique produit des chiffres proportionnels à cette quantité d'information.

Ces expérimentations ont été menées sur le corpus TimeBank [126] de textes anglais annotés en événements et en relations temporelles, mais également sur un jeu de données artificielles. Ce dernier permet d'avoir une quantité de données bien plus importante et

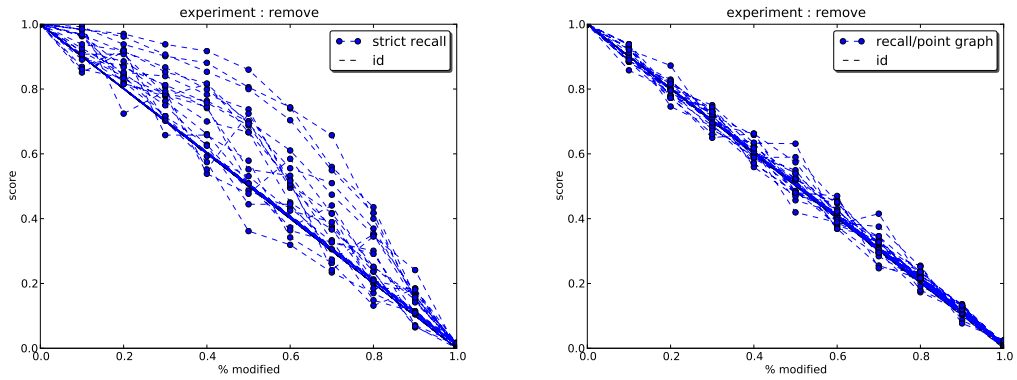

FIGURE 2.15: Comportement de la métrique proposée sur des graphes de 30 événements, rappel classique (à gauche) et rappel sur les points (à droite). En abscisse, le pourcentage de relations supprimées du graphe initial. En ordonnée, la valeur fournie par la mesure.

de contrôler les paramètres pouvant influencer les résultats, en particulier :

- Taille des graphes ;
- Indétermination des relations (du graphe complètement ordonnancé jusqu'à des configurations avec très peu d'informations temporelles).

Pour chaque graphe temporel (correspondant à un document textuel annoté), on réduit progressivement l'information temporelle en supprimant aléatoirement des relations, et on calcule les mesures à chaque étape. Cette opération est reproduite de nombreuses fois pour obtenir un nombre de situations statistiquement significatif. Un exemple de résultat pour le rappel est proposé à la figure 2.15, sur des graphes artificiels de 30 événements. On y voit que le rappel classique a un comportement très variable au fil de la suppression des relations, formant parfois une parabole très marquée. À l'opposé, notre mesure diminue de façon linéaire et avec de faibles variations.

2.5 Conclusion

Nous avons ici exploré l'extraction d'information temporelles et événementielles dans les textes, avec une vision relativement classique de l'extraction d'information et de relations, telle que présentée dans la première partie du chapitre 1 (section 1.1). Nous nous sommes intéressé aux désignations des événements et aux relations temporelles qui les lient, que ce soit dans le domaine général ou le domaine médical, avec des techniques symboliques ou par apprentissage. Nous avons également accordé l'importance nécessaire à l'évaluation, qui pose des problèmes originaux dans ces situations.

Les travaux de ce type ont des limites qui nous avons déjà citées. Ils semblent en particulier condamnés à se cantonner strictement à l'étude des phénomènes se produisant à l'intérieur d'un même document. Difficile en effet, lorsque toute l'approche est basée sur l'événement en tant que mot, de relier tel mot d'un document à tel autre d'un autre document, sans se heurter à des problèmes encore insolubles de référence, de contexte sémantique, de cohérence. Tout au plus peut-on espérer remplir des bases d'événements connus pour améliorer l'analyse des textes mentionnant ces événements, au prix d'un taux d'erreurs qui semble pour l'instant difficile à surmonter.

Les mérites du regain d'intérêt porté à ces tâches depuis quelques années sont pour-

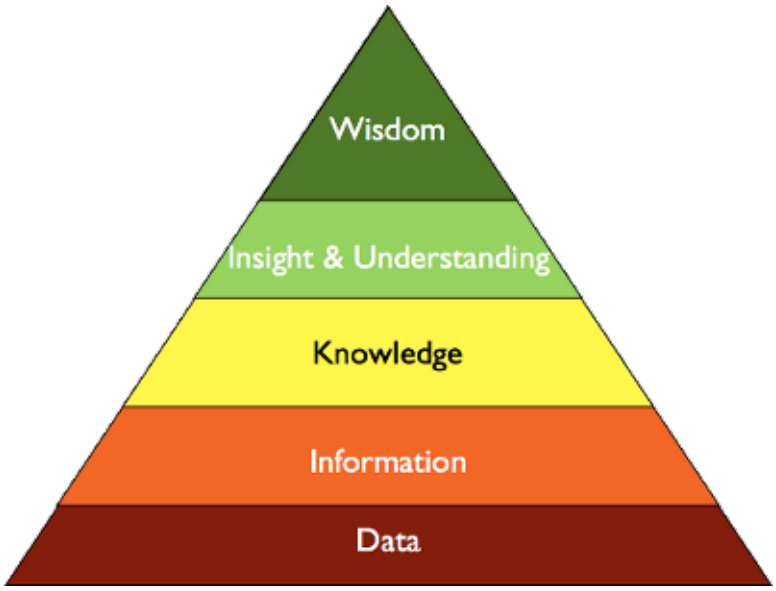

FIGURE 2.16: En route vers la sagesse ? La pyramide DIKW (*Data*, *Information*, *Knowledge*, *Wisdom*).

tant nombreux. Tout d'abord, ils ont permis un approfondissement des réflexions sur la notion d'événement en général, avec un travail de formalisation important, notamment dans le cadre de la norme TimeML. Ensuite, ils ont conduit au développement de nombreux outils d'analyse temporelle, dont certains auxquels nous avons participé, qui peuvent maintenant être utilisés aussi simplement qu'un analyseur syntaxique (avec d'ailleurs un taux d'erreurs comparable...) et appliqués à des tâches s'écartant autant que l'imagination le permet de celles des campagnes pour lesquelles ils ont été conçus. Le chapitre suivant en trahira d'ailleurs une utilisation intensive. Enfin, ils permettent d'entrevoir certaines applications originales, mentionnées également au chapitre 1, comme l'aide à la décision ou la fouille d'information guidée par les événements.

Pour notre part, tout en conservant un pied ancré dans ces travaux (certains étant très récents), nous avons plongé l'autre dans un milieu différent, moins fréquenté et moins bien balisé, dans lequel nous tentons de faire cohabiter l'extraction d'information fine avec des techniques de recherche d'information qui ont justement pour réputation d'empêcher toute finesse ; dans lequel l'information que nous extrayons est issue de l'analyse de millions de documents ; dans lequel nous tentons de faire émerger de la connaissance à partir du bruit ambiant ; dans lequel nous finissons même par essayer de transformer de l'information en donnée, alors que nous enseignons depuis des années que la donnée est inférieure à l'information dans notre quête de sagesse [133]. Tout ceci est le sujet du chapitre suivant.

Chapitre 3

L'événement dans le monde

Quand trop d'information sublime l'information

Les approches de l'événement que nous avons décrites au chapitre précédent sont très fortement associées aux segments textuels dans lesquels ces événements sont décrits. On peut dire qu'il s'agit d'une vision linguistique de l'événement. Nous nous en écartons maintenant pour aborder ce que nous appelons "l'événement dans le monde", et que nous pourrions qualifier de vision "journalistique" : quelque chose qui se passe, dont l'importance pour une certaine communauté fait qu'il est digne d'être mentionné, et qui nécessite un décryptage, par exemple en termes de causes, déroulement, conséquences.

Les technologies de l'information et de la communication ont fourni des outils et des méthodes pour démocratiser la production d'information. Une grande quantité de contenus de toutes sortes se déverse sur nous chaque jour, et le risque est fort que la connaissance que le citoyen pourrait acquérir reste au bout du compte enfouie sous le bruit. En l'absence d'organisation hiérarchique et de contextualisation, nous pouvons en effet manquer de perspective pour comprendre et assimiler la multiplicité des événements dont nous sommes informés chaque jour, et pour les relier à des événements passés. Nous proposons dans ce chapitre des stratégies pour utiliser la quantité d'informations disponibles en la présentant au lecteur de façon synthétique et structurée.

Nous suivons dans ce chapitre une progression comparable à celle adoptée dans la présentation du chapitre 1, tout en étant complémentaire, puisque les travaux décrits visent justement à combler certaines des lacunes observées dans ce même chapitre. Nous présentons donc d'abord une technique d'élaboration automatique de chronologies thématiques, dans lesquelles les événements sont décrits exclusivement par du texte [87, 88] (section 3.1). Nous nous penchons ensuite sur la question des relations entre événements dans un environnement [146] multidocument (section 3.2). Enfin, nous abordons une approche beaucoup plus orientée "données", avec un système d'analyse et d'aggrégation des relations d'alliances ou d'opposition entre pays (section 3.3).

3.1 Génération de chronologies thématiques

Une chronologie événementielle thématique est une liste de descriptions d'événements précis, associés à leur date d'occurrence, et considérés comme importants du point de vue d'un thème particulier (par exemple, "le Printemps arabe", "Michael Jackson", "les attentats à Bagdad").

Nous avons proposé une méthodologie de construction automatique de chronologies événementielles thématiques en français et en anglais, s'appuyant à la fois sur des techniques fines d'extraction d'informations, temporelles en particulier, et sur des opérations statistiques sur les masses de données, combinées par des techniques d'apprentissage [87, 88] [1]. Cette tâche, qui implique l'extraction des événements importants, est à mi-chemin entre la production de résumés multidocuments [112, 101] et les tâches de *Retrospective Event Detection* [161] ou *New Event Detection*, telles que définies dans les campagnes *Topic Detection and Tracking* (TDT) [5].

Les différentes éditions de la tâche TDT ont permis le développement de différents systèmes qui détectent la nouveauté dans les fils d'actualité [5, 94, 62]. La plupart de ces systèmes s'appuient sur des modèles statistiques à base de sac de mots utilisant des mesures de similarité pour déterminer la proximité entre les documents [99, 27]. Des efforts ont également été faits pour construire automatiquement des chronologies textuelles et graphiques [7, 37]. La majorité des systèmes conçus pour la construction de chronologies événementielles utilisent des approches par sac de mots, et très peu d'informations temporelles : généralement, seules les métadonnées du document, comme sa date de création, sont utilisées. Les quelques systèmes qui utilisent des informations temporelles n'extraient que des dates absolues (celles dont la lecture hors contexte suffit à les placer sur l'axe des temps, comme "14 juillet 1789" , contrairement à "14 juillet").

Notre travail se distingue des recherches précédentes dans la mesure où l'extraction d'événements importants est guidée par le traitement temporel. En effet, nous considérons que les événements importants, ceux que nous souhaitons retrouver dans les chronologies, ont lieu à des dates que nous pouvons également juger comme importantes – du point de vue du thème imposé par la requête de l'utilisateur. D'autre part, l'expression de ces événements dans les textes sera souvent accompagnée d'une expression temporelle, et répétée plusieurs fois si l'événement a une certaine importance – d'un point de vue journalistique.

Pour extraire ces dates importantes, la première étape est la collecte d'un maximum d'informations temporelles dans les textes. Ensuite, nous créons la notion de *date saillante*, définie à l'aide de critères essentiellement statistiques, avec pour but de nous rapprocher des dates présentées dans les chronologies de référence.

Ainsi que nous l'avons dit plus haut, dans la mesure où les dates apparaissant dans les textes viennent ancrer temporellement l'expression linguistique des événements, si une date est considérée comme saillante pour une thématique donnée, alors nous supposons que les événements auxquels ces dates sont rattachées sont certainement importants d'un point de vue informatif pour cette thématique. Elles seront donc de bonnes candidates pour la constitution des chronologies événementielles.

3.1.1 Description des ressources et du système

Nos chronologies de référence, produites manuellement par l'AFP, sont sous forme d'une liste de dates (généralement entre 10 et 20) associées à un texte décrivant l'événement ayant eu lieu à cette date. La figure 3.1 présente un exemple de chronologie produite par l'AFP. D'autres exemples sont donnés dans la figure 3.2.

La figure 3.14 présente l'architecture générale de notre système. Dans un premier temps, un prétraitement du corpus AFP permet d'annoter et de normaliser les expressions temporelles de chaque article (étape ① sur la figure). Le corpus ainsi enrichi est indexé par le moteur de recherche Lucene [2] (étape ②). L'ensemble des informations

1. Travail réalisé en collaboration avec Rémy Kessler, Caroline Hagège et Véronique Moriceau dans le cadre du projet ANR Chronolines.
2. http://lucene.apache.org

```
<NewsML Version="1.2">
      <NewsItem xml:lang="en">
          <HeadLine>Key dates in Thailand's political crisis</HeadLine>
          <DateId>20100513T100519Z</DateId>
          <NameLabel>Thailand-politics</NameLabel>
          <DataContent>
              <p>The following is a timeline of events since the protests began, soon after Thailand's Supreme Court confiscated 1.4 billion dollars of Thaksin's wealth for abuse of power.</p>
              <p>March 14 : Tens of thousands of Red Shirts demonstrate in the capital calling for Abhisit's government to step down, saying it is elitist and undemocratic. The premier and key ministers hole up in an army barracks.</p>
              <p>March 28 : The government and the Reds enter into talks but hit a stalemate after two days, as Abhisit refuses to meet a 15-day deadline for polls.</p>
              <p>April 3 : Tens of thousands of protesters move from Bangkok's historic district into the city's commercial heart, raising the stakes in the standoff.</p>
              <p>April 7 : Abhisit declares state of emergency in capital after Red Shirts storm parliament.</p>
              <p>April 8 : Authorities announce arrest warrants for protest leaders.</p>
              ...
          </DataContent>
      </NewsItem>
</NewsML>
```

FIGURE 3.1: Exemple de chronologie manuelle de l'AFP

- *Chronologies de 18 mois de troubles en Côte d'Ivoire*
- *Histoire des prises d'otages par des rebelles tchetchènes*
- *Les désordres politiques en Irak depuis l'élection du 7 mars*
- *Athlétisme : les records mondiaux du 800m*
- *Les accidents majeurs dans les mines chinoises*
- *L'espace en 2005 : un calendrier*
- *L'impasse sur le nucléaire iranien*
- *Chronologie de la guerre du Vietnam*
- *Chronologie de l'affaire Dominique Strauss-Kahn*
- *Le déroulement des attentats dans les transports londoniens*

FIGURE 3.2: Exemples de thèmes de chronologies manuelles de l'AFP

temporelles extraites est de cette façon associé au contenu textuel lors du processus d'indexation.

Étant donnée une requête, un certain nombre de documents est renvoyé par Lucene (③). Ces documents peuvent être filtrés (④), et les dates sont extraites des documents restants. Chaque date est ainsi accompagnée d'un ensemble de phrases pertinentes pour la requête donnée. Ces dates sont ensuite classées afin de montrer les plus importantes à l'utilisateur (⑤) accompagnées des phrases qui les contiennent. Toutes ces étapes sont détaillées dans les sections suivantes.

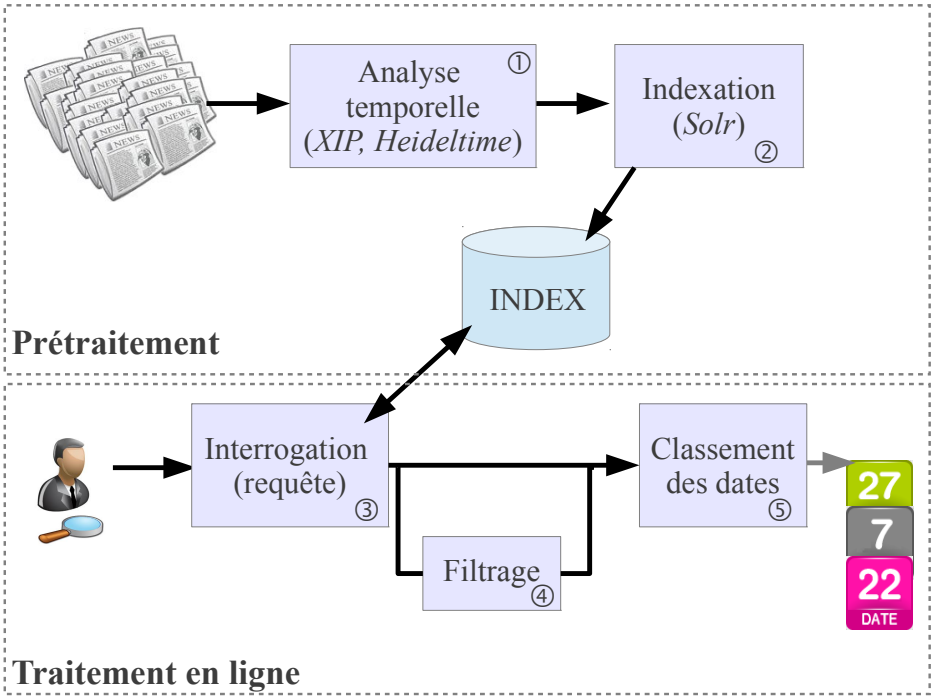

FIGURE 3.3: Présentation du système

3.1.2 Analyse temporelle des textes

Nous nous appuyons sur l'analyseur XIP [4] ou sur l'outil d'analyse temporelle Heideltime [143], que nous avons adapté à la langue française [118].

Le but de l'analyse temporelle est d'être capable de reconnaître, d'annoter et de normaliser [1] un sous-ensemble d'expressions temporelles que nous considérons pertinentes pour notre tâche.

Les deux outils cités opèrent le même type de normalisation des dates :

- Les dates *absolues* sont des dates qui peuvent être normalisées sans connaissance externe ou contextuelle. C'est le cas, par exemple, de l'expression *"le 5 janvier 2003"* . Dans ce type d'expressions, toutes les informations nécessaires à la normalisation sont contenues dans l'expression linguistique. Cependant, les dates absolues sont relativement peu fréquentes dans notre corpus (7 % de toutes les dates extraites en anglais, 12 % en français). Il est donc indispensable de considérer les dates relatives, beaucoup plus fréquentes, afin d'élargir la couverture pour la détection des dates saillantes.
- Les dates *relatives* à la DCT (date de création du document) représentent 40 % des dates extraites du corpus AFP en anglais et en français. Contrairement aux dates absolues, leur seule considération ne permet pas de procéder à leur normalisation. D'autres informations, en particulier la date correspondant au moment d'énonciation, mais également le temps des verbes rattachées aux expressions temporelles, sont nécessaires.

1. Nous appelons *normalisation* l'opération qui consiste à transformer une expression temporelle en une représentation formatée et entièrement spécifiée (par exemple, trouver la valeur absolue qui correspond à une date relative).

En première approximation, dans la mesure où nous disposons d'un large corpus dans lequel la redondance existe, nous avons estimé que nous devions à tout prix privilégier la précision et non le rappel. En effet, nous pensons que l'introduction de bruit dans la détection des dates saillantes serait plus préjudiciable que l'omission de certaines dates. Nous avons ainsi volontairement ignoré les cas "difficiles" comme ceux qui nécessiteraient la résolution d'anaphores entre éventualités, et nous avons considéré uniquement les dates non récurrentes, absolues ou relatives par rapport au moment de l'énonciation, qui est donné comme métadonnée et correspond à la date de création de la dépêche.

Dans le corpus AFP en anglais, comprenant environ 1,5 million de documents au moment de l'étude en 2011, 11,5 millions d'expressions temporelles ont été détectées par XIP, parmi lesquelles 845 000 sont des dates absolues (7 %) et 4,6 millions sont des dates relatives (40 %). En français, 9,4 millions d'expressions temporelles comprenaient 1,1 million de dates absolues (12 %) et 3,8 millions de dates relatives (40 %).

3.1.3 Expériences sur les dates saillantes et résultats

Dans toutes les expériences qui suivent, nous utilisons deux valeurs différentes pour classer les dates par ordre d'importance :

- Lucene fournit des documents classés associés à leur score de pertinence. $luc(d)$ est la somme des scores Lucene pour les unités textuelles contenant la date d ;
- une adaptation du *tf.idf* pour les dates :

$$tf.idf(d) = luc(d).log\frac{N}{df(d)}$$

où N est le nombre de paragraphes indexés et $df(d)$ est le nombre de paragraphes contenant la date d. Les $tf.idf$ sont sommés pour donner le score de la date.

Dans tous les cas, comme expliqué plus haut, la sortie du système est une liste de dates, ordonnées de la plus saillante à la moins saillante, accompagnées par des phrases associées à ces dates et au thème de la recherche. Un extrait simplifié d'une telle sortie est présenté à la figure 3.4.

Pour évaluer nos résultats, nous les avons comparés aux chronologies manuelles de l'AFP en utilisant la *Mean Average Precision* (MAP), qui est une métrique répandue et adaptée à l'évaluation de listes ordonnées. La MAP donne un poids plus fort aux éléments bien classés. Il s'agit de calculer la précision à chaque position de la liste ordonnée de résultats (ici, de dates) et d'en faire la somme tant que toutes les dates n'ont pas été trouvées. Pour une requête, la précision moyenne AP (*Average Precision*) est donc :

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{r}$$

où k est le rang de la date dans la liste, n le nombre de dates retournées, r le nombre de dates de la chronologie de référence, $P(k)$ la précision à ce rang et $rel(k)$ la valeur de pertinence (0 ou 1) de cette date. La MAP est la moyenne des précisions moyennes sur l'ensemble des requêtes.

Extraction de dates saillantes

Baseline

Dans la baseline que nous proposons, l'indexation et la recherche de document sont effectuées au niveau du paragraphe par le moteur de recherche Lucene (le titre et les

```
<data title="La crise politique à Magagascar" keywords="madagascar"
datemin="2009-01-16" datemax="2010-08-13">
    <event date="2009-03-17" rank="1">
        <doc date="2009-08-08">Lâché par l'armée, Ravalomanana avait remis le
        17 mars ses pouvoirs à un directoire militaire qui les avait immédiatement
        transférés à Andry Rajoelina, devenu depuis le nouvel homme fort de Ma-
        dagascar.</doc>
        <doc date="2010-10-11">Madagascar est plongée dans une grave crise po-
        litique depuis fin 2008 et l'éviction de M. Ravalomanana le 17 mars 2009,
        au profit de M. Rajoelina, ex-maire d'Antananarivo</doc>
        <doc date="2009-03-19">Elle avait auparavant validé l'ordonnance par la-
        quelle le président Ravalomanana, lâché par l'armée, a démissionné mardi
        en transférant les pleins pouvoirs à un directoire militaire, qui les a ensuite
        remis au chef de l'opposition.</doc>
        ...
    </event>
    <event date="2009-08-09" rank="2">
        <doc date="2009-08-09">Le sommet sur Madagascar organisé à Maputo
        s'est terminé dimanche, et la médiation a donné rendez-vous aux leaders
        malgaches dans une dizaine de jours au Mozambique, selon des média-
        teurs.</doc>
        <doc date="2009-08-12">La présidence de l'Union européenne s'est félici-
        tée dans un communiqué d'un accord signé dimanche à Maputo, qui prévoit
        la mise en place à Madagascar d'un gouvernement de transition avant des
        élections d'ici fin 2010.</doc>
        ...
    </event>
    ...
</data>
```

FIGURE 3.4: Exemple de sortie (simplifiée) du système de sélection des dates saillantes.

Modèles	Scores MAP	
	Corpus anglais	Corpus français
Baselines avec uniquement les dates absolues		
BL^{luc}	0,2782	0,3516
$BL^{tf.idf}$	0,2778	0,3528

TABLE 3.1: Scores MAP pour le système *baseline*

mots-clés du document sont ajoutés à chaque paragraphe). Pour une requête donnée, les 10 000 premiers documents sont renvoyés [1]. Seules les **dates absolues** sont prises en compte. Les dates sont classées selon une des deux valeurs décrites précédemment (luc or $tf.idf$). Nous avons ainsi obtenu les *runs* BL^{luc} et BL^{tfidf}.

Utilisation de la redondance uniquement

Dans ces expériences, l'index Lucene a été construit de la façon suivante : chaque **paragraphe** contenant une date normalisée est considéré comme un document. Le titre ainsi que les mots-clés de la dépêche AFP sont ajoutés à chaque paragraphe du

1. Ce chiffre a été défini de façon empirique.

document lors de l'indexation. Pour une requête, les 10 000 premiers documents sont renvoyés.

Cette série d'expériences est appelée par la suite *SD* [1]. Les dates obtenues sont classées en faisant la somme des scores Lucene des paragraphes associés (*luc*) ou par le $tf.idf$.

Utilisation d'autres informations, avec apprentissage

Nous avons utilisé l'ensemble de 91 chronologies manuelles comme corpus d'apprentissage afin d'entraîner nos modèles.

Nous utilisons IcsiBoost [2], une version libre du classifieur BoosTexter, fondé sur AdaBoost [61], un algorithme de *boosting* de classifieurs simples (des arbres de décision à 1 niveau de profondeur). Concernant le paramètre définissant le nombre d'itérations de l'algorithme, les expériences ont montré qu'un optimum était atteint aux alentours des 1 000 tours.

Dans notre approche, nous considérons deux classes : les dates saillantes sont les dates contenues dans les chronologies manuelles tandis que toutes les autres dates sont considérées comme des dates non saillantes. Cette approche présente cependant un biais important. Les choix des journalistes sont en effet très subjectifs et les chronologies ne devant pas dépasser une certaine longueur, des dates pertinentes peuvent être exclues.

Nous avons choisi de ne pas représenter chaque paragraphe comme une instance du classifieur mais plutôt de regrouper tous les paragraphes correspondant à la même date avant la phase d'apprentissage. Ainsi, chaque instance correspond à une date unique, et les traits associés concernent l'ensemble des paragraphes contenant cette date.

Les traits utilisés pour ces expériences sont les suivants :

1. traits représentant le fait que plus une date est mentionnée, plus celle-ci doit être importante :
   - somme des scores Lucene des paragraphes contenant cette date,
   - nombre de paragraphes contenant cette date,
   - ratio entre les scores Lucene de cette date et les scores Lucene de toutes les dates,
   - ratio entre le nombre de phrases contenant cette date et le nombre de phrases de toutes les dates ;
2. traits représentant le fait qu'un événement important est toujours mentionné, longtemps après qu'il se soit produit :
   - distance en nombre de jours entre la date et l'évocation la plus récente de cette date (la plus tardive dans le corpus),
   - distance en nombre de jours entre cette date et la DCT ;
3. autres traits :
   - meilleur classement Lucene de la date,
   - nombre de fois où la date est absolue dans les paragraphes (*i.e.*, nombre de paragraphes renvoyés, dans lesquels la date était une date absolue, normalisée sans recours à une connaissance externe ou contextuelle, par opposition aux dates relatives),
   - nombre de fois où la date est relative et normalisée dans les paragraphes,
   - nombre total de mots-clés de la requête présents dans le titre, le texte ou les entités nommées des documents renvoyés,
   - nombre de fois où la date est extraite d'un discours rapporté ou au futur.

1. Pour Salient Dates en anglais.
2. http://code.google.com/p/icsiboost/

Modèles	Scores MAP	
	Anglais	Français
Baseline (dates absolues seulement)		
BL^{luc}	0,2782	0,3516
$BL^{tf.idf}$	0,2778	0,3528
Redondance (et dates normalisées)		
SD^{luc}	0,6962	0,6722
$SD^{tf.idf}$	0,6982	0,6756
Tests avec apprentissage		
ML^{luc}_{base}	0,7033	0,6684
ML^{luc}	0,7905 **	0,6988
$ML^{tf.idf}$	0,7918 **	0,7008

TABLE 3.2: Scores MAP obtenus en combinant extraction de dates saillantes et filtrage simple. La signification statistique des résultats avec filtrage comparés aux résultats sans filtrage est indiquée par le t-test de Student (* : $p < 0,05$ (significatif) ; ** : $p < 0,01$ (très significatif)). L'utilisation du $tf.idf(d)$ par rapport au $occ(d)$ présente également une progression très significative dans les deux langues.

Nous n'avons pas pour but de classifier de manière binaire les dates, mais plutôt de les classer. Pour cette raison, nous avons utilisé la probabilité $P(d)$ renvoyée par le classifieur combinée au score Lucene des paragraphes ou au *tf.idf* de la date d :

$$score(d) = P(d) \times val(d)$$

où $val(d)$ est soit $luc(d)$ ou $tf.idf(d)$.

Nous avons évalué cette approche avec une validation croisée classique sur quatre partitions du corpus d'apprentissage. Chaque chronologie manuelle est ainsi répartie aléatoirement dans une des partitions. Nous utilisons trois sous-corpus en tant que données d'apprentissage (deux sont utilisés en apprentissage tandis que le dernier permet de calibrer le classifieur) et le quatrième corpus est réservé au test : quatre jeux de test tournants sont ainsi créés. Le score final d'un test est ensuite calculé en faisant la moyenne de tous les scores obtenus sur chacun des quatre corpus de test.

Résultats

Le tableau 3.2 présente les scores MAP obtenus par l'ensemble des approches décrites ici. Nous y voyons que la normalisation des dates relatives apportent des progrès importants, et que l'ajout d'autres critères que la redondance, par apprentissage, améliore encore les résultats.

Par la suite, une évaluation manuelle auprès de journalistes de l'AFP a confirmé que les chronologies produites étaient satisfaisantes.

Ces travaux sont les premiers que nous avons menés en nous intéressant directement à la notion d'importance et de classement des événements. Le restant de ce chapitre

reprend cette notion avec des intentions liées mais différentes, à commencer par une approche à la fois relationnelle et multidocument de cette importance.

3.2 Relations entre événements

Dans le chapitre 1, nous mentionnons que l'étude des relations entre les événements s'est surtout appliquée à l'intérieur d'un document, avec une vision de l'événement centrée sur le mot. Des relations temporelles entre des événements sont des liens entre des mots du texte. À l'exception de quelques travaux se situant dans la lignée des campagnes Topic Detection and Tracking (TDT [5]) [121, 54], la vision multidocument de l'événement n'a pas donné lieu à des travaux sur les relations.

Nous essayons ici de détecter les relations temporelles entre les événements tout en adoptant la vision de l'"événement dans le monde" que nous traitons dans ce chapitre [146] [1]. Nous utilisons la même collection de documents de l'AFP que dans la section précédente, en filtrant certains articles étiquetés comme des chronologies, des agendas (annonces des articles à venir) ou des notices biographiques ou descriptives. Nous partons ensuite du principe qu'un article est associé à un événement. En effet, même si plusieurs événements sont bien entendu mentionnés dans un article, nous supposons qu'un événement unique a déclenché la parution de cet article. Nous nous attachons ensuite à construire un graphe temporel d'articles (et donc d'événements), liés les uns aux autres par les relations suivantes :

- *Same event*, lorsque deux documents relatent le même événement, ou lorsqu'un document est une simple mise à jour d'un autre.
- *Continuation*, lorsqu'un événement est la "continuation" d'un autre (conséquence, suite naturelle, nouveaux développements...)

Nous définissons également un sous-ensemble de *continuation*, appelé *reaction*, représentant l'expression par quelqu'un d'un sentiment ou d'une opinion au sujet d'un événement.

Des exemples de ces trois classes de relations sont donnés à la section 3.2.1. Ces relations sont représentées par un graphe orienté où les nœuds sont les documents et les arcs sont les relations, comme le montrent les figures 3.5 et 3.6.

Comme nous allons le décrire après la description des données utilisées (section 3.2.1), ici encore, la redondance des descriptions d'événements va nous aider à être plus précis dans nos prédictions (section 3.2.2).

3.2.1 Ressources

Nous utilisons le corpus de l'AFP déjà présenté à la section précédente.

Sélection des paires d'articles

Des paires de documents ont été automatiquement sélectionnées pour annotation, en fonction des contraintes suivantes :

- La distance entre les dates de parution des deux articles ne doit pas dépasser 7 jours ;
- Au moins deux mots-clés ou deux noms propres du premier paragraphe sont en commun.

Il s'agit de restrictions importantes mais nécessaires, pour collecter assez de relations (un hasard total aurait peu de chances de conduire à des articles liés), tout en

1. Travail effectué avec Véronique Moriceau.

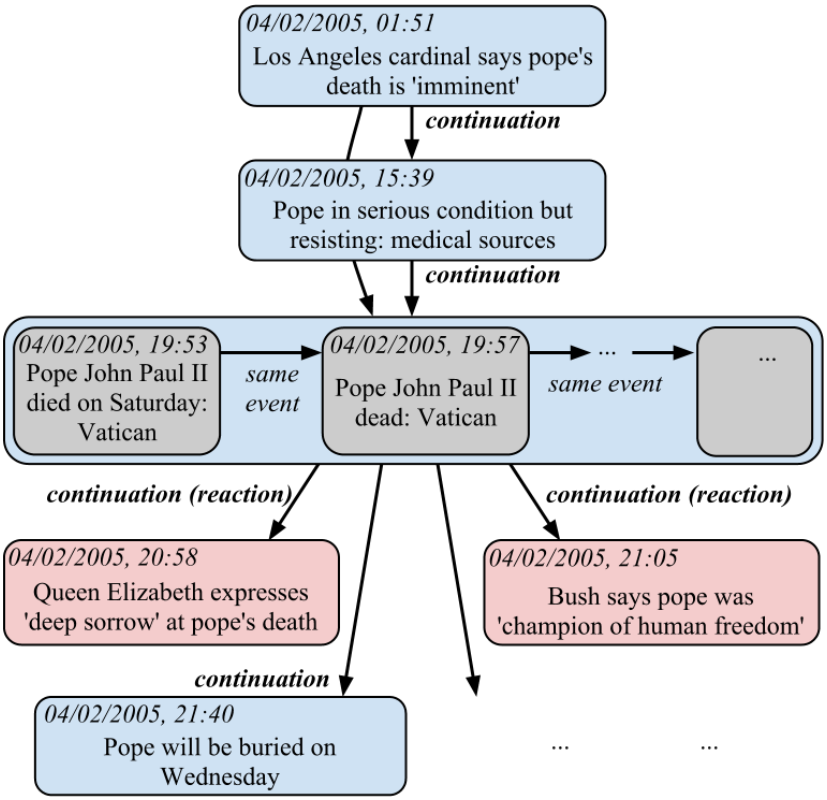

FIGURE 3.5: Un exemple de "graphe temporel" autour de la mort du pape Jean-Paul II. Le texte associé est le titre de chaque article. Les relations que l'on peut obtenir par transitivité sont masquées par souci de clarté..

garantissant des temps de calcul raisonnables pour l'application finale, toutes les paires d'articles devant être comparées.

Nous partons du principe que le titre et le premier paragraphe du document décrivent l'événement principal. Cette hypothèse est réaliste, une règle de base du journalisme imposant que le premier paragraphe, voire la première phrase, informe sur les "4 W" (*What*, *Who*, *When*, *Where*). Pourtant, lire davantage que le premier paragraphe est souvent nécessaire pour déterminer le lien entre deux articles. D'autre part, l'intégralité de l'article est d'une grande aide pour le système, puisque ce sont souvent les phrases suivantes qui assurent la redondance, en rappelant le contexte et les événements préalables.

Annotation des relations

Deux annotateurs ont attribué des relations à chaque paire d'articles présentée par une interface d'annotation dédiée.

Dans un premier temps, 7 types d'annotations étaient proposés :
- Trois relations autour de la notion de "même événement" :
  - *number update*, lorsqu'un nouvel article fournit une simple mise à jour d'informations numériques sur un événement (figure 3.9, en haut),
  - *form update*, quand le second document apporte de simples corrections mineures au premier,
  - *details*, quand le second document apporte plus de détails au sujet de l'événe-

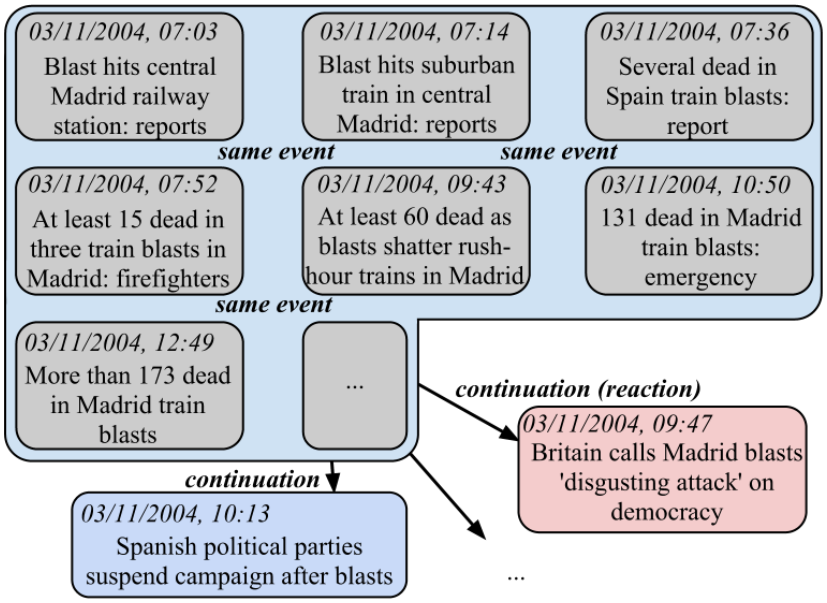

FIGURE 3.6: Un exemple de "graphe temporel" : les attentats de Madrid en 2004, avec de nombreuses mises à jour de l'information initiale. Notons que les articles regroupés dans le paquet principal (*same event*) peuvent avoir une date de parution postérieure à des articles de continuation ou de réaction.

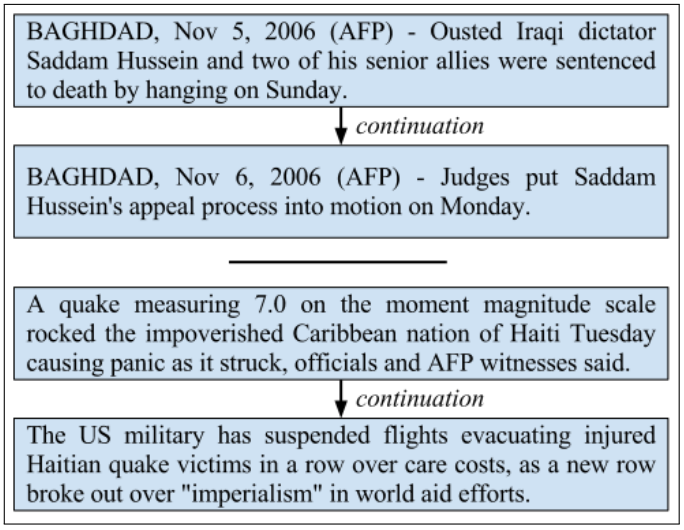

FIGURE 3.7: Exemples de la relation *continuation*.

ment (figure 3.9, bas).

- *development of same story*, lorsque deux documents mentionnent deux événements que l'on peut considérer comme inclus dans un troisième ;
- *continuation* (voir deux exemples à la figure 3.7). Cette relation est plus qu'une simple relation thématique, elle implique une prolongation naturelle entre deux événements. Par exemple, deux événements sportifs des mêmes Jeux Olympiques, ou deux différents attentats en Irak, ne seront pas reliées.
- *reaction*, un sous-ensemble de *continuation*, où un document relate la réaction de quelqu'un à un autre événement, comme illustré par l'exemple de la figure 3.8.

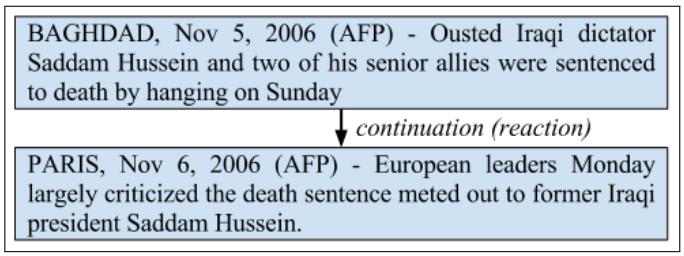

FIGURE 3.8: Exemples de la relation *continuation-reaction*.

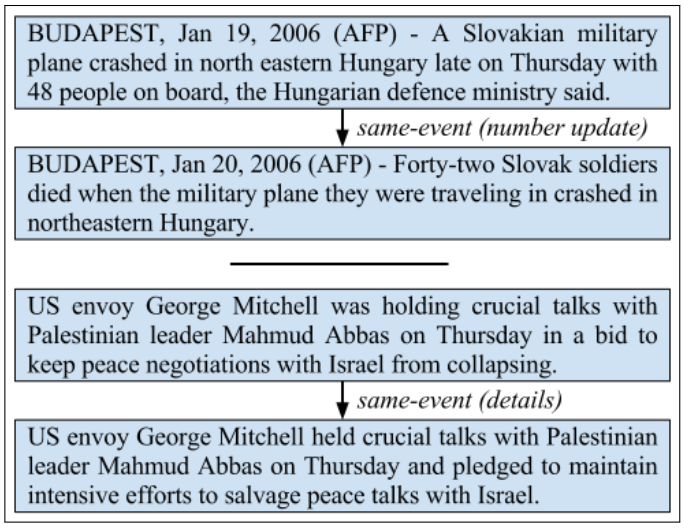

FIGURE 3.9: Exemples de relations *same-event* entre deux documents : mise à jour sur un nombre de victimes (en haut) ou plus de détails (en bas).

– *nil*, lorsqu'aucune relation n'est identifiée.

À l'issue de l'annotation de 150 paires, l'accord inter-annotateur s'est avéré bien trop faible (Kappa de Cohen [40] : $\kappa = 0,68$). Nous avons regroupé les relations *number update*, *form update* et *details* dans une relation générique plus consensuelle (*same-event*, voir la figure 3.9). Nous avons également supprimé la relation *development of same story*.

Après cette mise à jour du guide d'annotation et une nouvelle phase d'annotation, l'accord est monté à $\kappa = 0,83$, ce qui est considéré comme un bon accord.

Extension des relations

Dans le but d'accumuler plus de relations sans exagérer l'effort d'annotation manuelle nécessaire, nous avons utilisé les procédés suivants :

– Lorsque l'annotateur attribue une relation entre deux documents, le système d'annotation lui suggère d'autres paires à annoter contenant l'un ou l'autre de ces deux documents.
– Les relations *same-event* et *continuation* étant transitives [1], une fermeture transitive a été appliquée à la fin du processus, ce qui a permis de collecter de nouvelles

1. Si A *same-event* B et B *same-event* C, alors A *same-event* C, et respectivement pour la *continuation*.

	Nombre de paires	Apprentissage	Évaluation
Same event	762	458	304
Continuation	1134	748	386
Reaction	182	123	59
Nil	918	614	304
TOTAL	**2996**	1943	1053

TABLE 3.3: Caractéristiques du corpus.

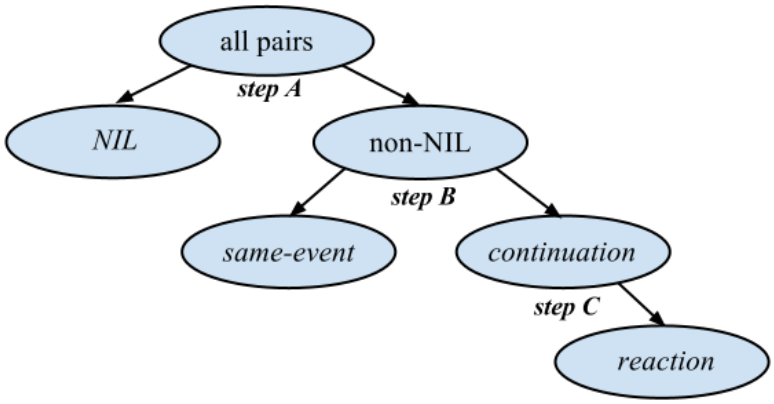

FIGURE 3.10: Classification en 3 étapes.

relations de façon automatique, mais également de détecter et de corriger quelques inconsistences dans l'annotation manuelle.

Au final, presque 3 000 relations ont été annotées, avec une répartition présentée à la table 3.3.

3.2.2 Construire les graphes temporels

De façon assez classique, nous avons séparé notre processus de classification en plusieurs étapes, chacune visant à distinguer une classe des autres, comme illustré à la figure 3.10 :

A. Classification entre *nil* et *non-nil* ;

B. Classification entre *same-event* et *continuation* ;

C. Extraction des relations *reactions*.

Dans cette section comme dans la suivante, nous utilisons le classifieur SMO [125, 86], optimisation des SVM, tel qu'implementé dans l'outil Weka [68], en utilisant les scores de prédictions (option "-M").

Nous n'entrerons pas ici dans le détail des classifieurs et des traits utilisés (qui pourront être trouvés dans [146]), mais nous décrivons tout de même deux aspects importants qui permettent de tirer partie de la redondance des informations et des caractéristiques particulières de nos relations.

En premier lieu, si un document est en lien avec beaucoup d'autres, on peut faire l'hypothèse que l'événement qu'il décrit est d'une certaine importance, soit parce qu'il est traité par plusieurs articles (relation *same-event*) soit parce qu'il a des conséquences (*continuation*). Ainsi, si un premier classifieur trouve de telles relations à propos d'un

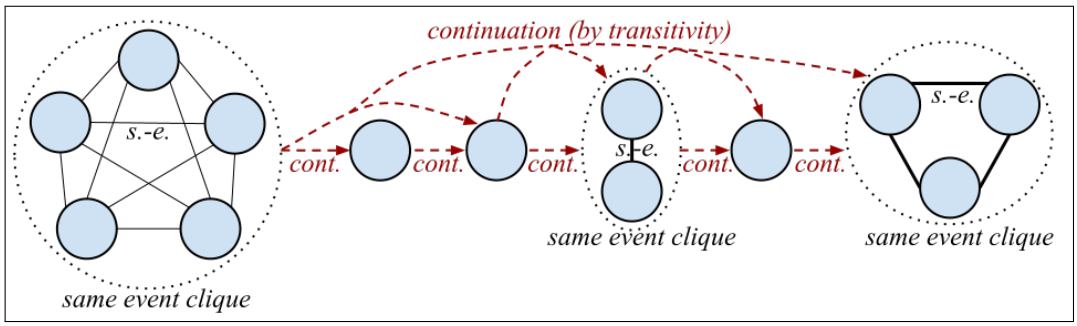

FIGURE 3.11: Un exemple de sous-graphe très connecté, correspondant au déroulement d'une séquence importante. Les articles relatant le même événement sont regroupés en cliques.

document, on peut être plus enclin à en trouver d'autres, puisque le document a de l'importance. Un exemple typique est présenté à la figure 3.11, où un événement décrit par plusieurs documents (à gauche) a de nombreuses continuations.

Pour tenir compte de cette intuition, à chaque étape A et B, la classification est elle-même divisée en deux phases, dont la deuxième utilise les résultats de la première, grâce à des traits supplémentaires concernant le nombre de relations entrantes ou sortantes trouvées pour chacun des deux documents de la paire concernée. Ceci permet d'améliorer de façon très significative les résultats [1].

Notre seconde façon d'exploiter les caractéristiques particulières de notre tâche est d'opérer une fermeture transitive du graphe par un mécanisme de vote. En effet, les relations *same-event* et *continuation* sont transitives, et*same-event* est également symétrique (A *same-event* B ⇒ B *same-event* A). Dans le graphe formé par les documents — nœuds — et les relations — arcs, il est alors possible de trouver toutes les cliques, c'est-à-dire les sous-ensemble de nœuds tels que toutes les paires de nœuds sont connectées par la relation *same-event*, comme l'illustre la figure 3.11.

À la fin de l'étape B, nous ajoutons donc une découverte des cliques de *same-event* avec l'algorithme de Bron et Kerbosch [29]. La fermeture transitive est ensuite telle qu'illustrée par la figure 3.12. Si le classifieur a proposé une relation entre des documents d'une clique et d'autres documents (par exemple, D1, D2 et D3), alors un vote est nécessaire :

- Si le document est lié à la moitié au plus de la clique par le classifieur, alors tous les liens manquants sont créés (figure 3.12.a) ;
- Dans le cas contraire, le document est totalement déconnecté de la clique (figure 3.12.b).

Ce vote est opéré pour les relations *same-event* et *continuation*. Notons qu'il n'apporte des améliorations que si la précision du classifieur initial est suffisante (selon nos expérimentations, au moins 70 %).

3.2.3 Application

L'application que nous proposons pour ce type d'approche peut se résumer au cas d'utilisation suivant :

1. Signification statistique mesurée par le test de Student.

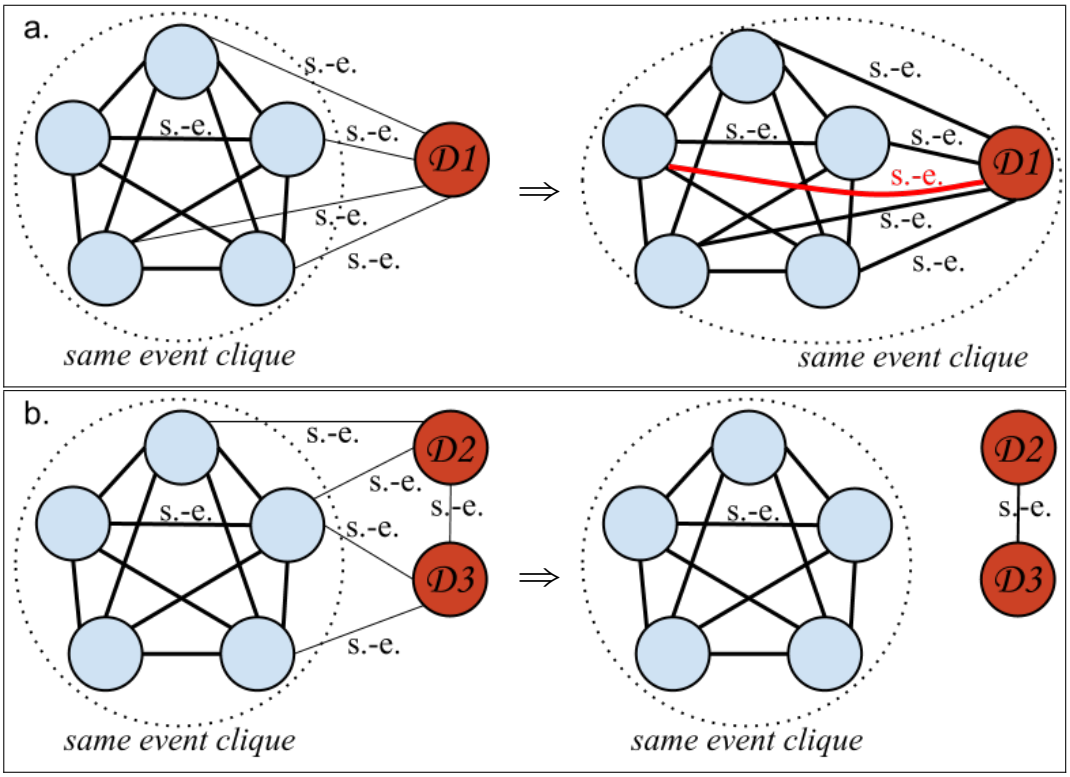

FIGURE 3.12: Vote pour la fermeture transitive des relations *same-event*. En haut (a.), quatre nœuds sur les cinq que compte la clique sont relié au document D_1 par le classifieur, ce qui est suffisant pour ajouter D_1 à la clique. En bas (b.), seulement deux nœuds de la clique sont reliés aux documents D_2 et D_3, ce qui est insuffisant. Tous les arcs entre la clique et D_2 or D_3 sont donc supprimés.

- Le lecteur lit un article (par exemple, sur la mort du pape Jean-Paul II, publié le 4 février 2005 (TU) — voir la figure 3.5).
- Un lien dans la page lui suggère un fil d'événement autour de cet article.
- Tous les articles dans une fenêtre de 7 jours autour de cet événement, partageant au moins deux mots-clés avec le document courant, sont collectés, et toutes les paires sont fournies au système[1].
- Quand des cliques de relations *same-event* sont trouvées, seul l'article le plus long (généralement, le plus récent) de chaque clique est présenté à l'utilisateur, mais la date et l'heure sont au contraire celles du plus ancien article relatant cet événement.
- Ceci permet de générer un nouveau graphe ne contenant que des relations de *continuation*. Les arcs sont "nettoyés", de façon à ce qu'un seul fil soit visible : les relations pouvant être obtenues par transitivité sont supprimées, et les arcs entre deux documents ne sont conservés que si aucun document ne peut être inséré entre les deux.
- Les nœuds sont présentés dans l'ordre chronologique. Ainsi, l'utilisateur peut visualiser et naviguer à travers le graphe (le fil ne montre que les titres mais les articles complets peut être accédés en cliquant sur le nœud).

1. Dans le cas d'événements trop importants où trop de paires seraient collectées, la fenêtre temporelle peut être restreinte.

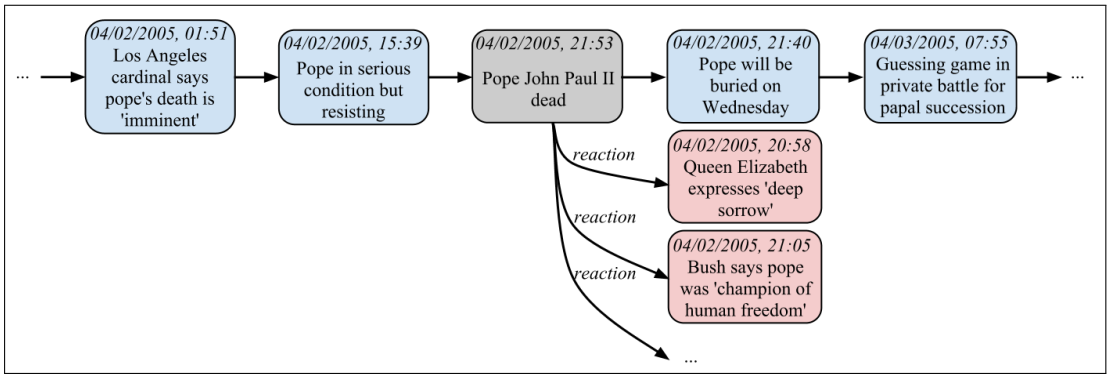

FIGURE 3.13: Un exemple de fil temporel d'événements, concernant la mort de Jean-Paul II (voir le graphe de relations correspodant à la figure 3.5).

– Lorsque des réactions sont trouvées, elles sont isolées du fil principal.
– Un tel fil temporel est potentiellement infini. Si l'utilisateur dépasse la fenêtre temporel de 7 jours en naviguant, le système est relancé.

La figure 3.13 présente le résultat de cette procédure sur le graphe temporel partiel de la figure 3.5.

Il existe clairement un lien entre le travail présenté à cette section et l'élaboration automatique des chronologies thématiques telle que proposée à la section 3.1. Dans les deux cas, une caractéristique essentielle de l'événement est son importance. De plus, chacun peut bénéficier à l'autre : une vision plus claire des relations entre les événements permet d'affiner une chronologie, et l'importance des dates telle que définie par notre travail sur les chronologies peut améliorer l'extraction des relations.

Nous quittons maintenant la vision exclusivement temporelle de l'événement que nous avons eue jusqu'à présent, pour étudier les relations entre entités qui sont introduites par les événements. Mais ceci est anecdotique, car l'apport principal du travail que nous allons présenter maintenant est la conversion d'informations extraites des textes en données statistiques, pour offrir une mise en perspective visuelle à un utilisateur, dans la droite ligne du journalisme de données.

3.3 Extraction et agrégation des alliances internationales

En effet, face à la quantité d'informations et à la multiplication des sources de données, journalistes et technologues ont développé la notion de "journalisme de données" [41, 51]. Cette pratique nouvelle du journalisme tire partie des données numériques disponibles pour produire et distribuer l'information. Elle bénéficie notamment de la popularité croissante de l'Open Data, du développement de bases de connaissances structurées comme DBPedia [98], YAGO [74] ou OpenCalais [1], ainsi que des travaux récents en visualisation de données, pour faciliter l'analyse de l'information et proposer une grande variété de points de vue.

Pourtant, la connaissance est encore loin d'être représentée entièrement dans des bases structurées, et la plus grande partie de l'information reste accessible uniquement sous forme textuelle. Le traitement automatique des langues a donc potentiellement beaucoup à offrir au journalisme moderne de ce point de vue.

1. www.opencalais.com

Nous avons tenté de montrer l'intérêt à la fois scientifique et applicatif de cette rencontre entre TAL et journalisme de données, en présentant un outil identifiant automatiquement les relations d'alliance et d'opposition entre les pays, sur un sujet spécifique défini par la requête d'un utilisateur (par exemple, la situation en Syrie, la prolifération nucléaire, la propriété du Pôle Nord). L'évolution de ces relations dans le temps — sur une échelle journalière, hebdomadaire, mensuelle ou plus — est alors illustrée par une courbe (voir plus loin la figure 3.17) ou par des graphes ou des cartes dynamiques (figures 3.19, 3.20 et 3.21).

3.3.1 Présentation du système

Nous appliquons des techniques classiques d'extraction d'information à une grande quantité de documents textuels, dans le but d'acquérir assez de données pour en déduire des statistiques significatives. Des données sont ensuite indexées par un moteur de recherche, puis utilisées par un outil de visualisation, par l'intermédiaire de requêtes.

Les relations que nous extrayons sont des relations d'opposition (*NEG*) ou d'alliance (*POS*) entre deux pays, décrites de façon explicite dans la même phrase. De telles relations sont illustrées par les exemples suivants.

(3.1) **Indonesia** voiced support for **East Timor**'s bid to join the ASEAN.
→ *POS(Indonesia, East Timor)*

(3.2) **China** earlier protested **Obama**'s meeting with the Dalai Lama, [...].
→ *NEG(China, U.S.A.)*

(3.3) **London**'s recent condemnations of **Libyan leader Moamer Kadhafi**'s bloody crackdown [...].
→ *NEG(U.K., Libya)*

(3.4) **Chavez** has stoop up for his longtime ally **Kadhafi**, [...].
→ *POS(Venezuela, Libya)*

L'alliance ou l'opposition peut principalement être rendue explicite par des verbes d'action (*protested*), des noms d'événement (*condemnations*) ou d'autres noms (*ally*). D'autre part, les pays (les arguments des relations) peuvent être désignés par leur nom, mais aussi par le nom de leur capitale ou par une personne représentant le pays. Dans nos exemples, il s'agit de présidents de pays, mais cela peut être bien plus large selon le contexte (par exemple, le nom d'un sportif). Plus de détails sur la classification des relations sont proposés à la section 3.3.2.

La figure 3.14 montre l'architecture générale du système. Lors du prétraitement, le corpus est analysé (étape ① et figure 3.15) par une segmentation en *chunks*, une résolution de coréférence, un étiquetage en entités nommées, la normalisation des informations temporelles et une association entre des noms de personnes et leurs pays.

Un classifieur extrait les relations *POS* et *NEG* dans chaque phrase. Ces informations sont ensuite indexées par le moteur de recherche Solr [93] [1] (étape ②), au niveau de la phrase.

Le traitement en ligne commence par l'arrivée d'une requête. Solr renvoie toutes les phrases à la fois pertinentes pour la requête et contenant une relation (③). Pour chaque jour du calendrier, les relations *POS* and *NEG* sont agrégées et une tendance de la journée est obtenue, ainsi qu'une évolution dans le temps (④). Enfin, une visualisation graphique de cette évolution est proposée à l'utilisateur (⑤ et figures 3.17, 3.19, 3.20 et 3.21).

1. http://lucene.apache.org/solr/

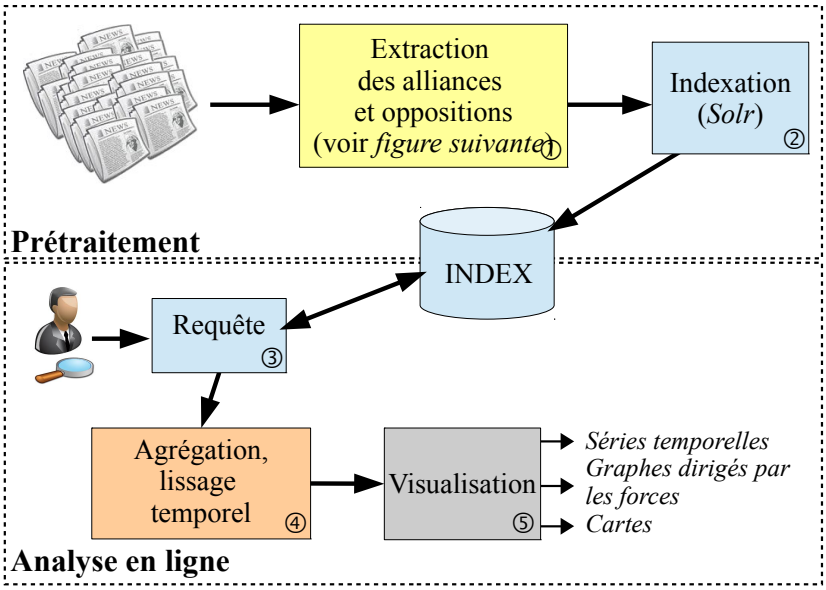

FIGURE 3.14: Architecture générale.

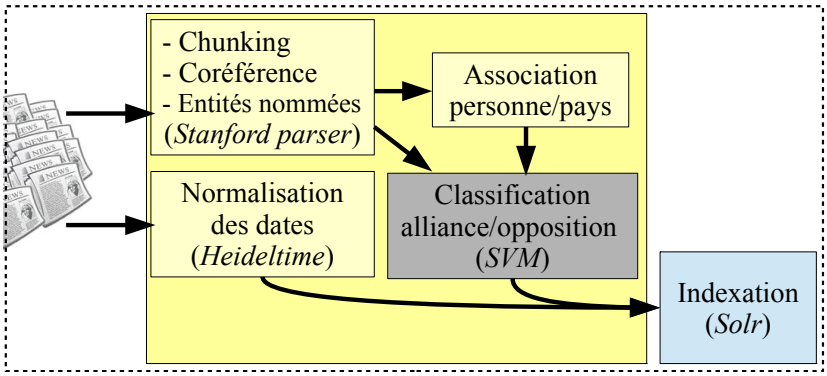

FIGURE 3.15: Détails des prétraitements effectués.

3.3.2 Classification entre alliance et opposition

Nous avons utilisé le corpus de l'AFP en anglais, déjà décrit plus haut, pour la phrase d'annotation et d'apprentissage, et nous avons ajouté un corpus d'environ un million de pages web de nouvelles, nettoyées avec Boilerpipe [91], pour l'application finale.

Entités.

Les exemples de la section 3.3.1 montrent que les relations entre pays peuvent être rendues explicites de plusieurs façons dans les textes. Les entités peuvent être des noms de pays, mais également des noms de ville ou de personne. Les noms de pays, les nationalités (démonymes) associées et les capitales sont extraites du "CIA World Factbook" [1], tandis que les variations linguistiques éventuelles de ces noms sont collectées grâce à DBPedia [98]. Nous tenons également compte des principales unions supranationalees (Nations Unies, Union Européenne, ASEAN, Union Africaine, etc.). Tous les éléments

1. http://wifo5-03.informatik.uni-mannheim.de/factbook/

cités sont considérés comme des entités, c'est-à-dire des arguments potentiels d'une relation d'alliance ou d'opposition.

Les noms de personnes demandent un traitement à part. Il s'avère en effet que les bases de connaissances existantes sont tout à fait incomplètes concernant l'association entre une personne et son pays. Nous avons donc utilisé une heuristique à la fois simple et efficace en examinant les cooccurrences des pays et des personnes dans les textes. Pour chaque nom de personne dans le corpus (identifié par le parseur de Stanford [90], avec la résolution de coréférences), le nom de pays (C) ou le démonyme $N(C)$ le plus proche dans la phrase est sélectionné. Un pays C est associé à un nom de personne P s'il représente plus de 30 % des cooccurrences entre P et l'ensemble des pays ou des noms de nationalité. Le nombre total des cooccurrences doivt également dépasser 100.

Nous avons évalué la précision de cette association en sélectionnant aléatoirement 100 noms de personnes respectant ces critères. 91 % de ces noms étaient associés au bon pays, et représentaient 96,1 % de la somme des occurrences des 100 noms dans le corpus. En d'autres termes, l'association entre un nom et un pays dans le corpus était un choix correct à 96,1 %. Sur les 9 erreurs, 8 étaient dues à des erreurs du parseur (les noms n'étaient en fait pas des noms de personnes).

Même si cette stratégie peut sembler simpliste, l'association est globablement correcte et concerne de nombreux sujets (diplomatie, politique, sport, problèmes militaires, etc.), ce qui rend le moteur de recherche applicable à tous les domaines.

Lexique.

Nos premiers tests pour l'identification des relations d'alliance et d'opposition s'appuyaient sur le lexique général "MPQA subjectivity lexicon" [160], contenant plusieurs milliers de mots annotés par leur polarité (positive ou négative). Ceci a cependant conduit à des résultats très bruités, et nous a persuadé de constituer notre propre lexique, bien plus restreint et basé sur une étude de corpus préalable sur un ensemble de développement. Ce lexique de "déclencheurs de relation" contient au final 110 mots utilisés pour exprimer l'alliance ou l'opposition, et non une notion de "polarité" bien trop générale. Ces mots sont des verbes (*agree*, *applaud*, *support*, *accuse*, *condemn*, *slam*...), des noms d'événement (*congratulation*, *accusation*, *sanction*...) et d'autres types de noms (*ally*, *criticism*...).

Normalisation des dates.

Pour normaliser les dates, et ainsi associer un marquage temporel à une relation, nous avons utilisé l'outil Heideltime [143], déjà décrit à la section 3.1.2.

Chunking.

Les arbres syntaxiques sont obtenus par l'application du parseur de Stanford. À partir de ces arbres, nous créons des *chunks* en regroupant les plus larges branches contenant des noms, des verbes ou des prépositions, mais jamais deux types différents d'éléments parmi ces trois. Par exemple :

(3.5) {In} {Jerusalem} {,} {Prime Minister Silvio Berlusconi} {pledged} {Italy's firm support} {for} {Israel} {and} {urged} {effective sanctions} {against} {Tehran}.

Classification

Au vu de la complexité des relations en jeu, il est important de noter que nous ne prétendons pas établir un classifieur générique des relations d'alliance et d'opposition. En effet, ces relations sont souvent non explicites dans les textes. Notre but est donc d'extraire assez de relations dans la collection, de façon à produire une quantité

de données significative, conduisant à une vision réaliste des relations internationales entre les pays. Au niveau de la classification, comme dans les sections précédentes, ceci se concrétise par un biais en faveur de la précision ; nous comptons ensuite sur la forte redondance des informations dans la collection pour obtenir une couverture satisfaisante des relations. Ainsi, le classifieur sera entraîné pour identifier des relations exprimées de façon explicite, dans la même phrase, en utilisant un déclencheur de la liste (section 3.3.2).

Ensemble d'apprentissage.
Nous avons sélectionné aléatoirement des phrases ayant les propriétés suivantes :

- Au moins deux entités (noms de pays, de capitale, d'union supranationale, de personne associée à un pays), séparées par moins de 15 chunks.
- Au moins un déclencheur de relation entre deux entités, ou moins de 4 chunks avant la première entité ou après la seconde.

Chaque paire d'entités satisfaisant ces contraintes est considérée comme une instance du classifieur. Une même phrase peut ainsi générer plusieurs instances. Ainsi, dans la phrase :

(3.6) In Jerusalem, Prime Minister **Silvio Berlusconi** pledged **Italy**'s firm *support* for **Israel** and urged effective *sanctions* against **Tehran**.

Les entités sont en gras et les déclencheurs de relation en italiques. Les instances de relations à classifier sont les paires suivantes :

- {Italy (Silvio Berlusconi), Israel}
- {Italy (Silvio Berlusconi), Iran (Tehran)}
- {Italy, Israel}
- {Italy, Iran (Tehran)}
- {Israel, Iran (Tehran)}

Les relations à trouver seraient *NEG(Italy, Iran)* et *POS(Italy, Iran)*.

Nous avons annoté 2015 instances avec la relation NIL (pas de relation, 1463 instances), NEG (opposition, 349 instances) et POS (alliance, 293 instances).

Classifieur.
Les phrases ne diffèrent pas dans leur structure, selon si la relation qu'elles expriment sont des alliances ou des oppositions. Les entités ont le même type d'interactions entre elles, et seule la polarité des déclencheurs compte. C'est pourquoi nous avons séparé le processus de classification en deux étapes, illustrée à la figure 3.16 :

A. Élimination des paires ne comportant aucune relation, c'est-à-dire classification entre *NIL* et *non-NIL* ;

B. Parmi les relations *non-NIL* relations, classification entre *POS* et *NEG* ;

Les classifieurs sont basés sur l'algorithme SMO [125, 86], une optimisation des SVM, tel que mise en œuvre dans le logiciel Weka [68], avec les paramètres par défaut. Les traits sont présentés à la table 3.4. Certains de ces traits ont été discrétisés, d'autres combinés entre eux.

Performances

Les résultats de ces classifieurs, évalués par validation croisée, sont présentés à la table 3.5.

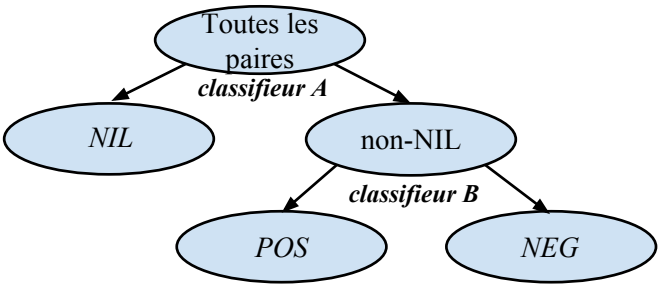

FIGURE 3.16: Une classification en deux étapes.

STEP A
Entité
- Distance entre les entités E_1 and E_2 - Longueur de la phrase - Positions de E_1 et E_2 dans la phrase - Type de E_1 et E_2 (capitales, nom de personnes, de pays) - Nombre d'entités entre E_1, à l'intérieur du chunk de E_1, entre E_1 et E_2, dans le chunk de E_2, après E_2
Traits lexicaux
- Présence d'un déclencheur de relation avant E_1, juste avant E_1 (chunk précédent), dans le chunk de E_1 ou E_2, juste après E_1, entre E_1 et E_2, juste avant E_2, juste après E_2, après E_2 - Nombre de déclencheurs à toutes les positions décrites à la ligne précédente - Nombre de déclencheurs à toutes les positions décrites à la ligne précédente, selon leur type (verbe, nom) - Négation entre E_1 et E_2 - Parmi les 20 prépositions les plus fréquentes dans le corpus : celles apparaissant juste avant et juste après E_1 ou E_2
Traits syntaxiques [1]
- E_1 (resp. E_2) est la tête de son chunk - Nombre de verbes avant E_1, entre E_1 et E_2 ; Nombre de verbes non-déclencheurs à ces positions - Nombre de noms avant E_1, entre E_1 et E_2 ; Nombre de noms non-déclencheurs à ces positions - Taille du chunk de E_1 (en mots), taille du chunk de E_2
Notons que des traits concernant le chemin syntaxique d'une entités à une autre ont été testé, sans succès.

STEP B
Distinction entre déclencheurs NEG et POS, négation
- Nombre de déclencheurs positifs (resp. négatifs) dans la phrase, ainsi que avant E_1, juste avant E_1 (chunk précédent), dans le chunk de E_1 or E_2, juste après E_1, entre E_1 et E_2, juste avant E_2, juste après E_2, après E_2 - Polarité globale (nombre de positifs - nomber de négatifs) à ces mêmes positions - Marques de négation à ces mêmes positions

TABLE 3.4: Traits utilisés pour les classifieurs A et B.

3.3.3 Agrégation temporelle et visualisation

Le classifieur décrit ci-dessus extrait un total de 330 222 instances de relations du corpus. Le rappel réel est plus faible que les chiffres donnés ci-dessus, et est difficile

Relation	Précision	Rappel
A, NIL vs. *non-NIL*		
non-NIL	**0,82**	0,76
NIL	0,90	0,93
moyenne	0,87	0,88
B, NEG vs. *POS*		
POS	0,95	0,99
NEG	0,99	0,95
moyenne	**0,97**	**0,97**
A + B, résultats globaux (POS & NEG)		
POS & NEG	**0,80**	0,73

TABLE 3.5: Validation croisée pour les classifieurs alliance/opposition.

à estimer, les données d'apprentissage étant limitées à certaines conditions (en particulier, la présence d'un déclencheur). Cependant, nous obtenons une relation tous les 5,4 documents en moyenne, ce qui semble être un ratio raisonnable, puisque la plupart des documents traitent de sujets tout autres que les relations d'alliance ou d'opposition entre les pays.

Marquage temporel des relations

La classification fonctionne au niveau de la phrase. Or, les événements décrits dans une phrase n'ont pas toujours lieu à la date de publication de l'article. De la même façon que pour la génération des chronologies (section 3.1), nous tentons d'améliorer le marquage temporel des relations, et donc la précision du système, en utilisant les dates normalisées par Heideltime. Si une date a été identifiée et normalisée dans la phrase, alors cette date marque la phrase. Sinon, la date de création du document est utilisée.

Indexation et interrogation

Un moteur de recherche Solr [93] est construit à partir des relations et de leur marquage temporel. Chaque phrase contenant au moins une relation est indexée avec son contenu, les pays concernés, les relations et la date associée. Une requête typique est ainsi composée des champs suivants :

keywords : quelques mots représentant le sujet ;
datemin, datemax : les bornes temporelles de l'étude ;
countries : zéro, une ou plusieurs noms de pays. Préciser des noms de pays signifie que l'on souhaite restreindre l'extraction des relations à ces pays. Le nombre de ces paramètres changera la façon de présenter les résultats (voir la section 3.3.3).

Agrégation de l'information

Toutes les phrases retournées par Solr, sans limite de nombre, sont agrégées. Pour chaque paire de pays considérée, dans un même jour d, si le nombre de relations *POS* et *NEG* entre les deux pays sont respectivement $P(d)$ et $N(d)$, alors le poids pour la paire et le jour d est :

$$w(d) = log(\frac{1 + P(d)}{1 + N(d)})$$

conduisant à une valeur comprise entre $-\infty$ et $+\infty$, où $w = 0$ signifie que la relation est neutre, $w < 0$ que la relation est une opposition, et $w > 0$ que la relation est une alliance.

Le bruit est alors réduit par un lissage, moyenne pondérée sur une fenêtre temporelle de W jours :

$$sw(d) = \frac{\sum_{j=-n}^{n} \frac{w(d+j)}{abs(j)+1}}{\sum_{j=-n}^{n} \frac{1}{abs(j)+1}}$$

où n est une demi-fenêtre ($n = floor(\frac{W}{2})$).

Par exemple, pour une fenêtre de 5 jours :

$$sw(d) = \frac{\frac{w(d-2)}{3} + \frac{w(d-1)}{2} + w(d) + \frac{w(d+1)}{2} + \frac{w(d+2)}{3}}{\frac{1}{3} + \frac{1}{2} + 1 + \frac{1}{2} + \frac{1}{3}}$$

Visualisation

Le mode de visualisation dépend de la requête de l'utilisateur. Si celui-ci a demandé des relations entre deux pays, alors la sortie sera une courbe de série temporelle. Dans les autres situations, nous obtiendrons un graphe basé sur les forces (*force-directed graph*) ou une carte interactive.

Série temporelle.

Pour les relations bilatérales (champ *countries* contenant deux éléments), un moyen évident d'informer les utilisateurs est de leur présenter une courbe représentant $sw(d)$, et de leur montrer sur demande les phrases qui conduisent à cette valeur. Un niveau de granularité (jour, semaine, mois) est alors précisé. Au niveau du jour, toutes les déclarations et réactions sont visibles, mais les données bruitées peuvent prendre une certaine importance. Des granularités plus larges conduisent à une courbe plus lissée et réduisent le bruit, ce qui peut être bon pour étudier les relations stables, à long terme.

La figure 3.17 montre un exemple de résultats concernant les relations entre les États-Unis et la Russie ("*countries : United States AND Russia*") au sujet de la situation en Syrie ("*keywords : syria*").

Graphe dirigé par les forces et carte.

Si zéro, un ou plus de deux pays sont spécifiés par l'utilisateur, une simple courbe ne suffit plus. Nous générons alors un graphe de pays, dans lequel la distance entre les nœuds reflète l'opposition entre les pays dans une période de temps données (plus large qu'un simple jour, cette fois). Le problème qui se pose ici est que le classifieur produit une relation entre deux pays de façon indépendante des autres relations ; il est donc très peu probable que les valeurs obtenues puissent être utilisées comme des mesures de distances entre des nœuds dans le graphe (en particulier, l'inégalité triangulaire ne sera pas respectée). Pour cette raison, nous construisons un graphe "dirigé par les forces" grâce à l'algorithme dit "*force-directed layout algorithm*" de Barnes-Hut [16], où une valeur entre deux nœuds est considérée comme une force de répulsion. L'algorithme essaie alors de minimiser l'énergie dépensée pour soutenir ces forces et maintenir le graphe relié, et donc d'adapter les distances entre les nœuds pour que le rendu sur un plan à deux dimensions soit possible.

Pour utiliser cet algorithme, nous devons transformer nos poids $sw(d)$ en nombres positifs (forces de répulsion). Nous devons également atténuer le bruit et les variations

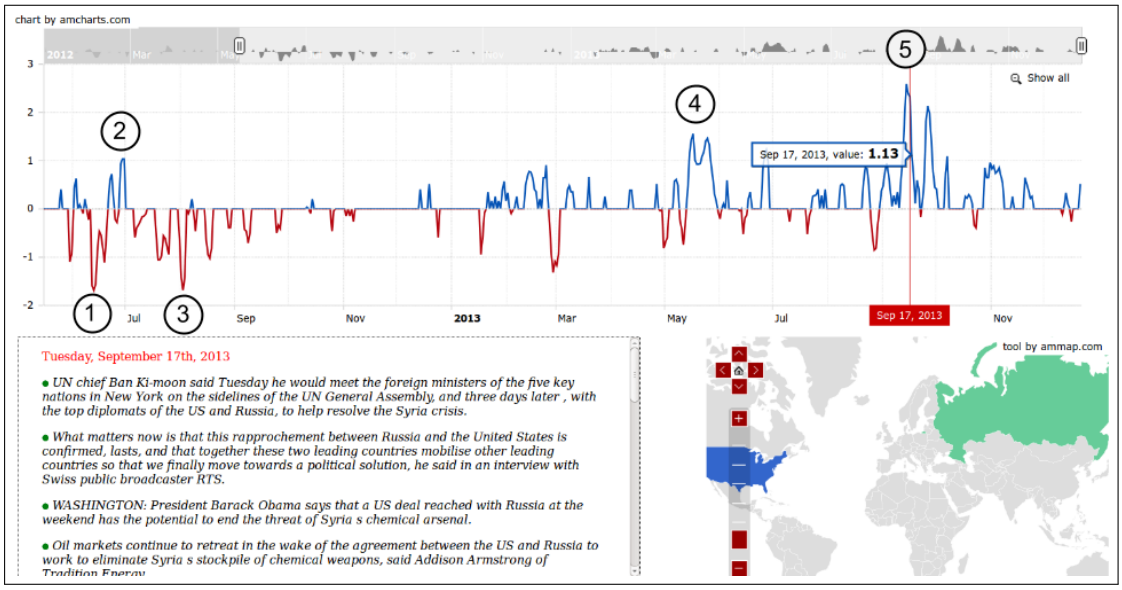

FIGURE 3.17: Exemple de résultat produit par le système pour les relations bilatérales entre les États-Unis et la Russie sur la requête "syria", à l'échelle du jour. Le cadre en bas à gauche montre les phrases correspondant à la date sélectionnée par l'utilisateur (ici, le 17 septembre 2013). Les nombres encerclés ont été ajoutés manuellement à la copie d'écran. Ils montrent l'évolution de ces relations, avec : ① Accusations mutuelles de fourniture d'armes aux autorités syriennes ou à l'opposition (relations d'opposition, $sw(d) << 0$) ; ② Prévision d'une réunion pour discuter de la situation (meilleures relations, $sw(d) > 0$) ; ③ Vetos de la Chine et de la Russie sur les résolutions des Nations Unies ; ④ Annonce d'une conférence de paix ; ⑤ Accord obtenu à cette conférence.

Relation	$sw(d)$	$sw'(d)$	Relation	$sw(d)$	$sw'(d)$
A_1 - A_2	1	0,27	B - A_1	-1	0,73
A_1 - A_3	2	0,12	B - A_2	-2,3	0,91
A_2 - A_3	3,5	0,03	B - A_3	-3	0,95

TABLE 3.6: Exemples de poids logarithmiques $sw(d)$ et de leurs ajustements logistiques $sw'(d)$.

de ces poids, inévitablement introduits par le volume de données. Ce problème est illustré à la table 3.6, où trois pays A_1, A_2 et A_3 sont alliés contre un autre (B) sur un thème donné et dans une certaine période de temps. Les poids $sw(d)$ des trois relations entre les alliés sont positifs mais varient entre 1 et 3,5, tandis que le poids de la relation B - A_1 est de -1. En conséquence, la paire A_1 - A_2 ($sw = 1$) est finalement plus proche de B - A_1 ($sw = -1$) que de A_2 - A_3 ($sw = 3,5$), ce qui n'est pas souhaitable.

Ce type d'effets peut être corrigé par une fonction logistique, "en forme de S", qui modélise un niveau de saturation après et avant une phase de croissance exponentielle (ou, dans notre cas, une décroissance) :

$$sw'(d) = 1 - \frac{1}{1 + e^{-sw(d)}} \tag{3.7}$$

Cette fonction, dont la forme est illustrée à la figure 3.18, atténue les poids élevés (valeurs négatives comme positives), accentue les différences entre les valeurs positives

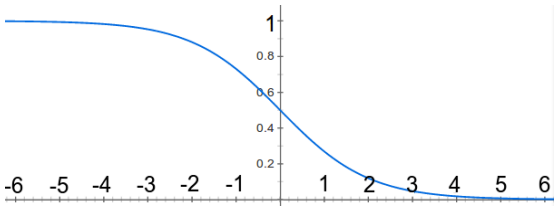

FIGURE 3.18: Fonction logistique (équation 3.7) utilisée pour transformer les poids des relations entre pays en valeurs utilisables comme des distances.

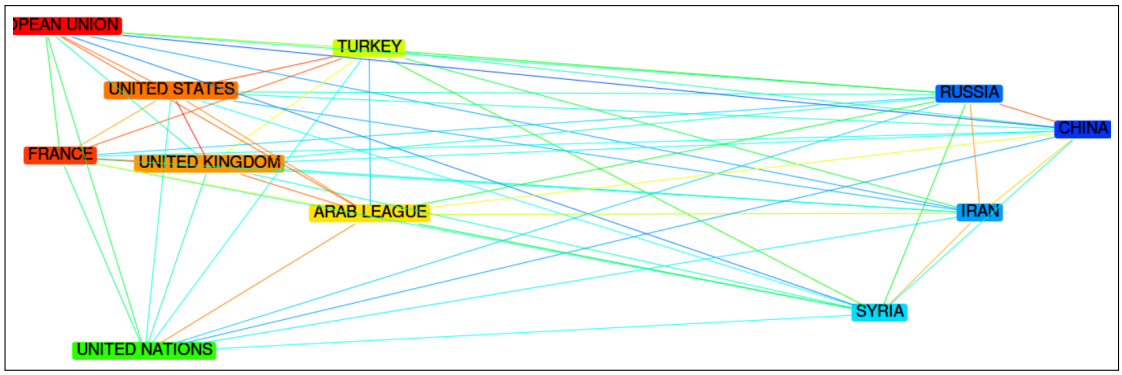

FIGURE 3.19: Exemple de graphe produit par le système pour les relations entre différents États sur la requête “syria”, pour l’année 2012. Le graphe est construit avec des informations collectées dans un total de 18 582 phrases. Les couleurs des arcs indiquent le type de relation (du rouge vif pour une alliance forte au bleu foncé pour une opposition forte). Les couleurs des nœuds réflètent la proximité des pays entre eux.

et négatives, et contribue ainsi à réduire le bruit sans avoir à discrétiser les valeurs de façon arbitraire. Les valeurs résultantes sont toutes positives, entre 0 et 1, ou 1 est une opposition forte (répulsion dans le graphe) et 0 une alliance forte (attraction dans le graphe), tandis que 0,5 est neutre. Les valeurs de sw' pour notre exemple sont également indiquée à la table 3.6, et l’on peut voir que l’inconvénient décrit ci-dessus est gommé.

Les figures 3.19 et 3.20 montrent deux exemples de tels graphes obtenus par la méthode décrite, avec dans la première une grande quantité de données collectées (*keywords : syria*) et dans la seconde des informations plus éparses (*keywords : austerity*).

Si imprimées en couleurs ou visualisées sur un écran, les figures montrent également que les couleurs des nœuds reflètent leur proximité entre eux, clarifiant les différents alliances au premier coup d’oeil. De la même façon, la seconde vue est une carte dynamique dans laquelle les pays concernés sont colorés de la même façon (figure 3.21).

Pour choisir ces couleurs, nous avons d’abord calculé le plus cours chemin visitant une fois l’ensemble des nœuds (en d’autres termes, il s’agit du fameux problème du voyageur de commerce), en partant du nœud le plus proche du coin supérieur gauche. Nous utilisons un simple algorithme choisissant à chaque étape le plus proche voisin. Cette méthode est loin de fournir la solution la plus exacte mais est rapide et simple à mettre en œuvre, tout en étant largement suffisante pour notre problème. Ensuite, le chemin est projeté sur une ligne en une dimension représentant une échelle de couleur s’étalant du rouge vif au bleu foncé.

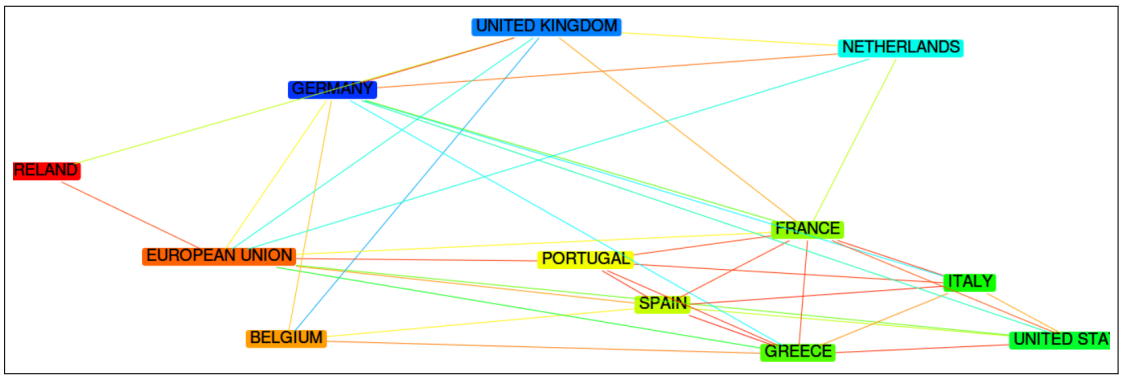

FIGURE 3.20: Exemple de graphe produit par le système pour les relations entre différents états sur la requête "austerity", entre 2012 et fin-2013. À cette époque, les pays du Sud de l'Europe s'opposaient à la politique d'austérité défendue par les pays du Nord. Le graphe est construit avec des informations collectées dans un total de 553 phrases.

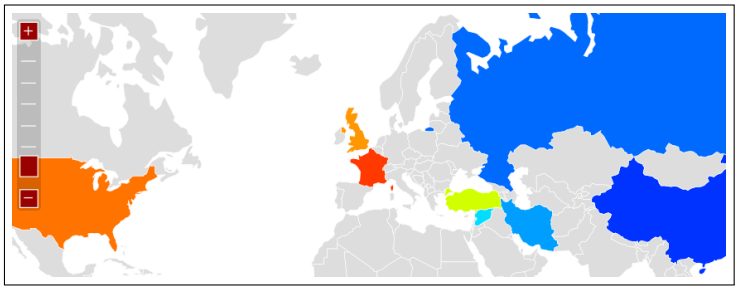

FIGURE 3.21: Exemple de carte produite, à partir des mêmes informations qu'à la figure 3.19.

3.3.4 Discussion

Comme pour toutes les approches d'extraction d'information, obtenir des données numériques précises et robustes avec des techniques de traitement automatique des langues dans de grandes quantités de données est une question d'équilibre entre précision et rappel. La spécificité se situe ici dans la manière de parvenir à cet équilibre. Comme dans d'autres applications comme les systèmes de questions-réponses sur le Web [52, 109, 106], les conseils à suivre sont les suivants :

- Penser d'abord à la précision. Ne certes pas négliger le rappel, car le but est d'obtenir des données, mais la grande quantité d'information disponible pourra compenser le manque de couverture [162].
- Contrairement à beaucoup de situations, une grande variation langagière est ici préférable. Les chances de collecter des données avec une classifieur précis seront plus fortes, de même que celles de réduire l'importance donnée aux erreurs de classification.
- Si cela a du sens, la dimension temporelle doit être considérée. En effet, certaines données ne sont valables que dans une période de temps limitée, et observer l'évolution des données est souvent d'un grand intérêt.

Par ailleurs, l'efficacité du système est tout à fait acceptable. L'analyse de la collecte dans son intégralité prend quelques jours sur un simple serveur à 32 cœurs, et seulement quelques minutes sont nécessaires pour analyser les nouveaux articles arrivant chaque

jour. Quant aux requêtes, les plus longues demandent quelques secondes.

Dans ce travail, l'innovation ne vient pas de nouveaux modèles, techniques ou représentations. En nous inspirant des travaux issus des campagnes Topic Detection and Tracking (TDT [5]), de l'analyse d'opinions [102, 150, 75] et des méthodes classiques d'extraction de relations [115, 83, 25, 165], nous montrons qu'il est possible d'agréger l'information et d'en déduire des données numériques fiables qui pourraient difficilement être obtenues avec d'autres moyens. Ceci permet d'apporter une valeur ajoutée significative aux journalistes et aux utilisateurs finaux.

3.4 Conclusion

Le système que nous venons de décrire nous permet de conclure ce chapitre consacré à nos explorations des facettes de l'extraction d'information dans des grands volumes de documents. Les progrès parallèles de la puissance de calcul d'une part, et des techniques d'extraction et de recherche d'information d'autre part, permettent en effet d'envisager d'allier analyse profonde et analyse statistique sur des grandes quantités de texte.

Si ce mouvement, que l'on peut considérer comme initié par les campagnes TDT, est déjà dynamique, les pistes à suivre restent nombreuses, que ce soit en termes d'applications ou de problèmes scientifiques. Il manque encore également de structuration et de formalisation pour ce qui est des méthodes et des représentations.

Comme nous l'avons vu, nous nous intéressons notamment aux liens avec le journalisme, non pas uniquement parce que nous manipulons des données journalistiques (articles de l'AFP ou pages web de sites journalistiques), mais parce que nous cherchons à contribuer à la pratique du journalisme, et à participer à son évolution. D'autre part, de façon beaucoup plus récente, avec des travaux encore trop peu avancés pour figurer dans ce mémoire, nous tentons d'appliquer la même vision de l'information au domaine bio-médical.

Nous aurons l'occasion de revenir sur tout cela dans le chapitre consacré aux perspectives (chapitre 6). Avant cela, nous dédions les deux prochains chapitres à des problèmes scientifiques sensiblement différents, que nous regroupons dans une partie intitulée "ciblage d'information". S'ils s'éloignent des sujets précédents par le fait qu'ils ne s'intéressent pas spécifiquement aux événements, ils conservent de nombreux points communs avec eux (ciblage, aspects multidocuments, association entre recherche et extraction d'information) et ont sans conteste nourri notre réflexion générale sur la place de l'information, sa représentation et ses liens potentiels avec le traitement automatique des langues.

Deuxième partie

Ciblage d'information

Chapitre 4

L'analyse syntaxique au secours des systèmes de réponse à des questions

Le contexte du travail présenté ici est celui des systèmes dits de "questions-réponses", Un tel système cherche à cibler l'information de façon précise, en réponse à la question d'un utilisateur, formulée en langage naturel. On peut ainsi, en particulier, répondre à des questions d'ordre factuel (*qui*, *quand*, *combien*, *où*, et leurs variantes) en fournissant directement l'entité réponse, ainsi qu'un texte "support", c'est-à-dire le ou les extraits de documents dans lesquels la réponse a été trouvée. Ce mécanisme s'oppose notamment aux moteurs de recherche traditionnels qui, à partir d'une requête formulée à base de mots-clés, renvoient des documents entiers à l'utilisateur.

Dans ce que nous présentons dans ce chapitre, l'objectif est de valider des réponses en vérifiant que toutes les informations données dans une question sont bien retrouvées dans les passages de texte supportant la réponse. Cette vérification repose sur le fait de retrouver les différentes entités précisées dans la question correctement reliées entre elles, soit dans une même phrase, un même passage, ou dans plusieurs documents.

Une information recherchée peut être présente sous différentes formulations et peut requérir, pour être reconnue, l'utilisation de bases de connaissances sémantiques et la réalisation d'inférences avec plusieurs pas de raisonnement pour relier les différentes parties de la réponse. Or, si l'on dispose bien de bases lexicales contenant des variations sur les termes (synonymes par exemple), des bases conceptuelles permettant de relier les concepts ne sont pas disponibles pour la plupart des langues et on ne peut donc envisager une analyse sémantique.

Aussi, notre propos consiste à retrouver les informations précisées par la question en se reposant sur la syntaxe, et notamment la reconnaissance des dépendances entre syntagmes et leurs différentes paraphrases possibles pour identifier les relations cherchées.

Nous présentons ici tout d'abord la stratégie générale mise en œuvre dans le système que nous avons appelé FIDJI (section 4.1), puis l'analyse syntaxique (section 4.2) et l'extraction et la justification de la réponse (section 4.3), avant de conclure avec une brève présentation des résultats (section 4.4) [1].

1. Travail réalisé avec Véronique Moriceau.

4.1 Présentation de FIDJI

La plupart des systèmes de questions-réponses peut extraire une réponse à une question factuelle quand celle-ci est explicitement présente dans le texte, mais ils ne sont pas capables de combiner plusieurs informations pour produire une réponse. FIDJI (Finding In Documents Justifications and Inferences), un système de questions-réponses en domaine ouvert pour le français et l'anglais, vise à introduire des mécanismes de compréhension reposant sur des inférences.

L'objectif est de produire des réponses qui sont entièrement validées par des extraits de textes (des passages).

La principale difficulté est qu'une réponse (ou des informations composant une réponse) peut être validée par plusieurs documents. Par exemple :
Question : Quel premier ministre français s'est suicidé ?
Réponse : Pierre Bérégovoy
Passage 1 : Le premier ministre français Pierre Bérégovoy a mis en garde Bill Clinton contre...
Passage 2 : Deux ans plus, Pierre Bérégovoy s'est suicidé après avoir été impliqué...

Dans cet exemple, la réponse *Pierre Bérégovoy* ne peut être entièrement validée par un seul passage : les informations *premier ministre français* et *s'est suicidé* sont validées par deux passages différents. En fait, la réponse est ici l'intersection de deux ensembles de réponses obtenues aux deux questions *Qui s'est suicidé ?* et *Qui sont les premiers ministres français ?* Dans ce cas, pour pouvoir proposer une réponse entièrement validée, il est nécessaire de décomposer la question en deux sous-questions.

Une analyse syntaxique peut fournir de telles décompositions pour les questions. Beaucoup de systèmes de questions-réponses utilisent des informations syntaxiques, en particulier les relations de dépendance, principalement pour l'extraction des réponses. Deux approches émergent : la première consiste à rechercher un appariement exact entre les relations de dépendance de la question et celles d'un passage [85], tandis que la seconde approche calcule une distance d'édition entre les arbres représentant la question et le passage [100].
[84] propose une stratégie pour décomposer les questions à un niveau syntaxique et sémantique : ceci permet à leur système de rechercher des informations dans plusieurs ressources. Il utilise un certain nombre d'"annotations paramétrées" et de patrons sémantiques appliqués à toute la collection de documents afin de relier une question aux informations d'un ou plusieurs documents. Ce système est principalement construit pour répondre aux questions portant sur des objets ou des propriétés (par exemple, une date de naissance, la population d'une ville, etc.). Le système de questions-réponses IRSAW pour l'allemand [70] adopte lui aussi une stratégie de décomposition des questions en s'appuyant sur un analyseur syntactico-sémantique.

Notre but est d'extraire et de valider des réponses en allant au-delà d'un appariement syntaxique exact entre la question et le passage, et cela sans utiliser de ressources sémantiques. Dans ce contexte de validation de réponse, nous avons remarqué que la stratégie de validation à appliquer (validation par un seul ou plusieurs documents) peut être guidée par la question, et en particulier par le type de réponse attendu. En effet, beaucoup de questions factuelles attendent une réponse d'un type qui peut être :

- une entité nommée comme dans *Qui a obtenu le Prix Nobel de la paix en 1995 ?* qui attend une réponse de type PERSONNE ;
- un type plus précis comme dans *Quel président russe a assisté au G7 en 2007 ?*, qui attend aussi une réponse de type PERSONNE mais dont le type est précisé explicitement dans la question (*président russe*). Le type précis n'est pas issu

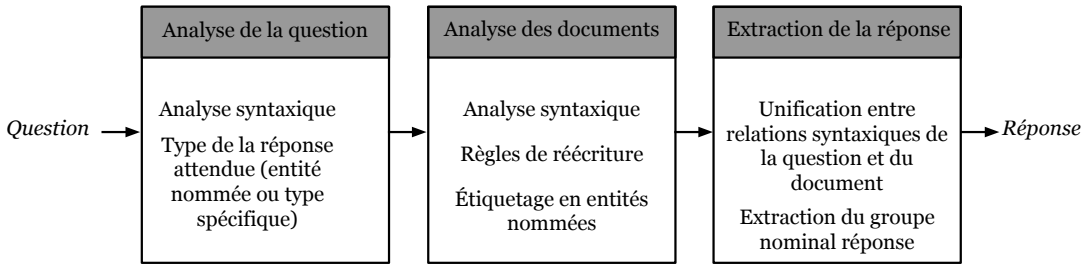

FIGURE 4.1: Architecture de FIDJI

d'une liste prédéfinie : il est identifié automatiquement lors de l'analyse de la question.

À la vue des différents exemples étudiés, nous faisons l'hypothèse que le type de la réponse est un élément qui peut être validé dans des documents autres que ceux d'où la réponse est effectivement extraite.

Ainsi, FIDJI utilise des informations syntaxiques, en particulier des relations de dépendances, qui vont permettre notamment de décomposer les questions. Le but est alors de vérifier la présence des relations syntaxiques de la question dans un passage justificatif et de confirmer le type de la réponse potentielle dans ce même passage ou dans un autre document, ceci afin de valider entièrement la réponse. La figure 4.1 présente l'architecture de FIDJI.

4.2 Analyse des questions et des documents

Notre système s'appuie sur les analyses syntaxiques produites par l'outil XIP [4], un analyseur robuste pour le français déjà présenté au chapitre 2 (section 2.1). XIP est utilisé pour analyser à la fois les questions et la collection de documents d'où sont extraites les réponses.

4.2.1 Analyse des questions

L'analyse des questions consiste à identifier :
- les dépendances syntaxiques : elles sont fournies par XIP ;
- le type de la question : factuelle, définition, booléenne, liste ;
- le type de réponse attendu : soit une entité nommée et/ou un type plus précis.

Les questions attendant des réponses de type liste sont celles qui contiennent explicitement un type de réponse au pluriel (par exemple *Quelles planètes... ?*, *Qui sont les... ?*, etc.).
Lors de l'analyse syntaxique de la question, la réponse à extraire est représentée par une variable (notée REPONSE) introduite dans les relations de dépendance. Le type de la question est principalement déterminé grâce à la forme de la question (pronom interrogatif, etc.). Le type spécifié de réponse attendu, s'il existe, est quant à lui le groupe nominal lié à la variable réponse par une relation *attribut* (équivalent à une relation *est_un*). Par exemple :

Q : Quand fut construite la Tour Eiffel ?
Dépendances : `DATE(REPONSE)`
`SUJ(construire, Tour Eiffel)`

```
AUX(construire, être)
```
Type de la question : factuelle
Type de réponse attendu : DATE (entité nommée)

Q : Quelle déclaration fut adoptée par l'ONU en 1948 ?
Dépendances :
```
attribut(REPONSE, déclaration)
AUX(être, adopter)
SUJ(adopter, REPONSE)
modif_par(adopter, ONU)
DATE(1948)
```
Type de la question : factuelle
Type de réponse attendu : déclaration (type plus précis)

Dans le premier exemple, le type de la réponse est une entité nommée alors que dans le second exemple, c'est un mot ou groupe de mots directement extrait de l'analyse syntaxique de la question.

4.2.2 Analyse des documents

Notre approche consiste à déterminer, pour une question donnée, si toutes les caractéristiques de la question (en l'occurrence les dépendances syntaxiques) peuvent être trouvées dans un ou plusieurs documents. Dans ce but, FIDJI détecte les implications syntaxiques entre la question et les passages. Toute la collection de documents est donc également analysée syntaxiquement par XIP.

Comme les informations dans les documents ne sont pas toujours exprimées de la même façon que dans les questions (par exemple, par le biais de variations syntaxiques), il est indispensable de raisonner sur les relations de dépendance syntaxique. De la même façon que [26], nous avons implémenté environ 40 règles de réécriture pour tenir compte, entre autres, des variations comme les changements voix passive/active, les nominalisations de verbes [77], les appositions, coordinations, etc. Ces règles de réécriture sont appliquées à toute la collection analysée.

De cette façon, quelle que soit la forme syntaxique de la question, le système est capable de trouver une formulation équivalente dans un passage justificatif : il est ainsi possible d'avoir un appariement exact soit entre les dépendances syntaxiques de la question et d'un passage, soit entre les dépendances syntaxiques de la question et celles d'un passage obtenues par réécriture.

L'exemple suivant illustre la règle de réécriture pour la reformulation passif/actif (les relations obtenues par réécritures sont en italique).

Q : Quelle ville a été secouée par un tremblement de terre le 17 janvier ?
```
DATE(17 janvier)
attribut(REPONSE, ville)
SUJ(secouer, REPONSE)
AUX(être, secouer)
modif_par(secouer, tremblement)
attribut_de(tremblement, terre)
```

Texte : Le tremblement de terre qui a secoué, lundi 17 janvier, Los Angeles ne serait pas associé directement à la fameuse faille de San-Andreas...
```
SUJ(secouer, tremblement)
```

```
OBJ(secouer, Los Angeles)
SUJ(secouer, Los Angeles)
AUX(être, secouer)
modif_par(secouer, tremblement)
attribut_de(tremblement, terre)
DATE(17 janvier) ...
```

Les relations de dépendance réécrites sont obtenues par application de la règle de la figure 4.2. Dans cet exemple, toutes les relations de la question sont présentes dans l'analyse du passage.

4.2.3 Autres ressources : les entités nommées

Les entités nommées des documents sont étiquetées en utilisant un ensemble d'environ 160 types [132] (*e.g.* personne, organisation, montagne, fleuve, nationalité, date, nombre, etc.). Cet étiquetage, associé aux résultats de l'analyse de la question, est utile pour vérifier que le type d'entité nommée attendu par la question est bien le même que celui de la réponse extraite dans le document. Par exemple, la question suivante attend une réponse de type LIEU :

Question : Où Barbara Hendricks a-t-elle donné son premier concert de l'année ?

```
LIEU(REPONSE)
SUJ(donner, Barbara Hendricks)
OBJ(donner, concert)
attribut_de(concert, année)
```

Passage étiqueté : <pers>Barbara Hendricks</pers> a donné son premier concert de l'Année nouvelle à <ville>Sarajevo</ville>.

```
SUJ(donner, Barbara Hendricks)
OBJ(donner, concert)
attribut_de(concert, année)
...
```

Dans cet exemple, toutes les relations de dépendance syntaxique de la question sont présentes dans le passage et le type de réponse attendu (LIEU) correspond bien au type de la réponse extraite *Sarajevo* (le type VILLE est un sous-type de LIEU dans la hiérarchie des entités nommées).

```
SUJ(Verbe, NP1)          SUJ(Verbe, NP2)
OBJ(Verbe, NP2)    ⇒     AUX(être, Verbe)
                         modif_par(Verbe, NP1)
```

FIGURE 4.2: Exemple de règle de réécriture : voix active vers voix passive.

4.3 Recherche de la réponse

4.3.1 Sélection des passages

Les documents sont indexés par le moteur de recherche Lucene [71]. Les mots-clés utilisés pour interroger l'index sont l'ensemble des mots significatifs de la question [1] ; les 100 premiers documents sont conservés (ainsi, seuls ces documents sont considérés par la suite).

4.3.2 Extraction de la réponse

La stratégie employée pour l'extraction de la réponse dépend du type de la question. L'utilisation de l'analyse syntaxique se situe au niveau de la phrase. Les informations syntaxiques, lorsqu'elles sont utiles, sont combinées avec d'autres paramètres.

Questions factuelles

Les phrases candidates sélectionnées sont celles qui comptent le plus de relations en commun avec la question. Une fois ces phrases obtenues, deux cas sont à considérer :

1. Au moins une dépendance de la question comportant la variable 'REPONSE' s'unifie avec une dépendance de la phrase. Dans ce cas, cette variable est instanciée par un lemme, qui est considéré comme la tête de la réponse. La réponse complète est ensuite composée en ajoutant les modifieurs présents dans la phrase (compléments du nom et adjectifs). Si cette réponse est du type d'entité nommée attendu, un poids plus important lui est attribué (voir la section 4.3.4).
2. Si la variable 'REPONSE' ne trouve pas d'instanciation dans la phrase, des éléments ayant le type d'entité nommée approprié sont recherchés. Ceci a pour but de contrebalancer les erreurs d'analyse. Il arrive bien entendu régulièrement que la réponse soit présente dans le passage mais que les seules relations syntaxiques n'y conduisent pas.

   Dans le cas d'absence d'entité nommée du type recherché dans la phrase considérée, elles peuvent être collectées dans les 2 phrases précédentes (avec un score moindre). Cette heuristique compense (très imparfaitement) l'absence d'une résolution anaphorique dans le système ; ceci produit inévitablement un certain bruit, mais qui est généralement masqué par la redondance des bonnes réponses (la fréquence d'extraction d'une réponse augmente en effet son score – section 4.3.4).

Questions listes

L'analyse de XIP, enrichie par certaines des règles de réécriture, fournit des informations sur la coordination des syntagmes.

Si une réponse est trouvée (par les techniques indiquées à la section précédente), les éventuelles relations de coordination avec cette réponse sont recherchées, et une liste est construite à partir des éléments ainsi collectés.

Si les réponses multiples sont préférées, les réponses atomiques sont retournées malgré tout ; en effet, une question étiquetée "liste" peut être satisfaite par un autre moyen. Par exemple, à la question *Qui sont les Dalton ?*, le pluriel indique certes que l'on attend une liste de noms, mais la réponse *"4 bandits"* est également tout à fait valable.

1. C'est-à-dire les mots étiquetés comme nom, verbe, adjectif ou adverbe par l'analyseur syntaxique.

Questions booléennes

Répondre à des questions booléennes est relativement proche de la tâche de validation de réponse, comme pratiquée par exemple lors des campagnes AVE (Answer Validation Exercise [130]). Dans cette tâche, les participants doivent considérer des questions, des réponses de systèmes et un texte de justification, et décider si la réponse à la question est à la fois correcte et validée entièrement par le texte.

Nous avons participé à AVE en langue française en 2008 : FIDJI s'est classé premier pour le français et second sur l'ensemble des candidats toutes langues confondues [119].

Nous utilisons la même technique pour les questions booléennes : si la part de dépendances de la question trouvées dans la phrase dépasse un certain seuil (déterminé empiriquement à 70 %), la réponse retournée est 'oui'.

4.3.3 Validation du type de la réponse

Lorsqu'une phrase est sélectionnée et une réponse extraite, il arrive souvent que toutes les relations recherchées ne soient pas satisfaites. Dans certains cas, il est alors possible de vérifier la présence des dépendances manquantes dans d'autres documents. Dans l'état actuel du système, la seule validation effectuée par ce biais est celle du type précis de la réponse.

Ce type est fourni, lorsqu'il existe, par l'analyse de la question (voir section 4.2.1). Si le passage justificatif ne contient pas les informations permettant de le vérifier, une nouvelle question est construite pour valider la réponse candidate.

Dans notre exemple *"Quel premier ministre s'est suicidé..."*, le type de réponse est *ministre* (tandis que le type d'entité nommée est PERSONNE). Le type *étendu* est *premier ministre*.

Q : Quel premier ministre s'est suicidé en 1993 ?
Dépendances : `SUJ(se suicider, REPONSE)`
`DATE(se suicider, 1993)`
`attribut(REPONSE, ministre)`
`attribut(ministre, premier)`

Si on retrouve dans un texte la phrase *Pierre Bérégovoy s'est suicidé en 1993*, la variable REPONSE s'unifie avec *Pierre Bérégovoy* qui devient une réponse candidate : les deux premières dépendances de la question sont ainsi vérifiées dans cette phrase. Il manque les deux suivantes, concernant le type précis (Pierre Bérégovoy était-il premier ministre ?).

La validation est opérée en deux étapes. Tout d'abord, le système vérifie que la réponse candidate est bien un *ministre*, en recherchant une relation nommée 'attribut' (`attribut(Bérégovoy, ministre)`). Si cela est confirmé, le type étendu est également vérifié et les deux relations `attribut(Bérégovoy, ministre)` et `attribut(français, ministre)` sont attendues dans la même phrase.

4.3.4 Justification et classement des réponses candidates

Classement des réponses

À chaque réponse est associé un quadruplet :
– la présence de tous les noms propres de la question dans la phrase (0 ou 1),
– la réponse a le bon type d'entité nommée (0 ou 1),
– la réponse a le bon type précis (0, 1 pour le type simple, 2 pour le type étendu),

Position de la réponse correcte	Rang 1	Rang 1 à 3
CLEF 2005	59,5 %	68,5 %
CLEF 2006	48,5 %	57 %

TABLE 4.1: Position des réponses correctes.

– le nombre de fois où la réponse a été extraite.

Les critères suivants sont utilisés pour classer les réponses (du plus important au moins important) :

1. Si le passage justificatif ne contient pas tous les noms propres de la question, alors la réponse est disqualifiée ;
2. Une réponse possédant le bon type d'entité nommée est préférée (quels que soient les scores ci-dessous) ;
3. Une réponse ayant validé le type précis est préférée également (le type étendu étant idéal) ;
4. Enfin, si tous les critères ci-dessus sont équivalents, une réponse trouvée plusieurs fois additionne les poids obtenus par chaque occurrence. La redondance est une information importante qui permet de masquer bon nombre d'erreurs.

Passage justificatif

Le passage justificatif est produit de la façon suivante :

– Pour chaque réponse, la phrase ayant le meilleur score est sélectionnée.
– Les phrases précédentes sont également incluses dans le passage (dans le but de collecter le contexte et d'éventuels antécédents anaphoriques) dans une limite de 256 caractères (limite classique des campagnes d'évaluation).

4.4 Évaluation et Conclusion

Nous avons évalué notre système sur les données de test des campagnes d'évaluation CLEF 2005 et 2006 [153, 105]. La collection de documents est composée d'environ 177000 articles de journaux en français (Le Monde et ATS 1994-1995 (environ 2 Go)). Ces documents sont supposés syntaxiquement corrects. Les questions de CLEF 2005 sont factuelles ou de type définition tandis que les questions de CLEF 2006 sont factuelles, de type définition ou liste.

Lors de ces campagnes, les sytèmes sont autorisés à proposer 3 réponses pour chaque question, ces réponses devant être classées par ordre de confiance.

Outre les performances du système en elles-mêmes (tableau 4.1), qui sont autour de l'état de l'art de l'époque (2008-2009), une évaluation intéressante est celle de la stratégie de décomposition des questions dans le but de valider une réponse grâce à plusieurs passages.

Dans l'ensemble de questions de CLEF 2005, il y a 51 questions (25 %) qui peuvent être décomposées en sous-questions en appliquant notre stratégie. FIDJI trouve une réponse correcte (sans tenir compte du rang) pour 32 questions (68,5 %). Chaque fois que la décomposition des questions est appliquée, le système peut rechercher des justifications à la réponse dans des documents différents.

Parmi les 32 réponses correctes pour les questions qui ont été décomposées en sous-questions, 22 % d'entre elles ont une justification dans plusieurs documents. Si FIDJI n'utilise pas la stratégie de décomposition, seulement 64 % des réponses sont correctes (au lieu de 68,5 %).

Dans l'ensemble de questions de CLEF 2006, 56 questions peuvent être décomposées. FIDJI trouve une réponse correcte pour 29 d'entre elles : 14 % de ces réponses correctes ont une justification dans plusieurs documents. Si FIDJI n'utilise pas la stratégie de décomposition, seulement 55 % des réponses sont correctes (au lieu de 57 %).

Nous avons montré dans ce travail comment l'utilisation de la syntaxe peut aider un système de questions-réponses à mieux cibler les résultats, notamment par une stratégie de décomposition de la question et une recherche multidocuments des justifications.

Chapitre 5

Collecte ciblée thématique de documents sur le Web

Nous choisissons de présenter ce dernier exemple de ciblage d'information car il se distingue clairement des autres par la taille de l'unité d'information ciblée. Ici, le but n'est plus de rechercher une information précise à l'intérieur d'un document, que cette information soit représentée par des mots (chapitres 2 et 4) ou des descriptions et représentations plus complexes (chapitre 3). Dans ce chapitre, nous nous rapprochons plutôt du domaine de la recherche d'information sur le Web, et plus précisément des moteurs de recherche dits "thématiques". Contrairement aux moteurs "généralistes", comme Google, Yahoo ou Bing, ceux-ci se concentrent sur un thème ou un domaine en particulier (la médecine, l'astronomie, la bourse, etc.). Ils permettent ainsi de mieux cibler l'information, au sens où il répondent à des besoins d'information plus précis en limitant le champ de leur index, ce qui aura pour effet de permettre par exemple l'utilisation de connaissances spécifiques au domaine, la mise en place d'une interface de recherche dédiée ou une réduction des ambiguïtés (dans un moteur spécialisé en médecine, le mot *virus* n'est plus ambigu).

Nous proposons dans ce chapitre un échantillon [48, 47] de nos recherches dans ce domaine [46, 49], à travers l'utilisation d'un moteur de recherche généraliste pour collecter le corpus spécialisé qui composera la collection de recherche du moteur thématique. À partir d'un petit ensemble de termes d'un domaine, servant d'amorces, nous appliquons une méthode de filtrage non supervisé à base de marche aléatoire dans un graphe hétérogène [1].

Notre objectif est de partitionner le corpus produit par un moteur généraliste et de ne conserver que le sous-ensemble traitant du thème voulu. Plutôt qu'effectuer un partitionnement strict des documents, nous proposons d'attribuer un poids à chaque document en fonction de son attachement au thème, ou plus précisément, aux termes amorces qui représentent notre thème. Un poids est également affecté à chaque terme, représentant sa proximité thématique aux autres termes amorces.

L'ordre produit pour les documents peut finalement être utilisé pour sélectionner les documents les plus pertinents, alors que l'ordre produit pour les termes peut permettre de sélectionner de nouveaux termes amorces pertinents en vue de soumettre de nouvelles requêtes.

Cette méthode du *bootstrapping* (amorçage) a déjà été utilisée dans la création de

1. Travail réalisé avec Clément de Groc.

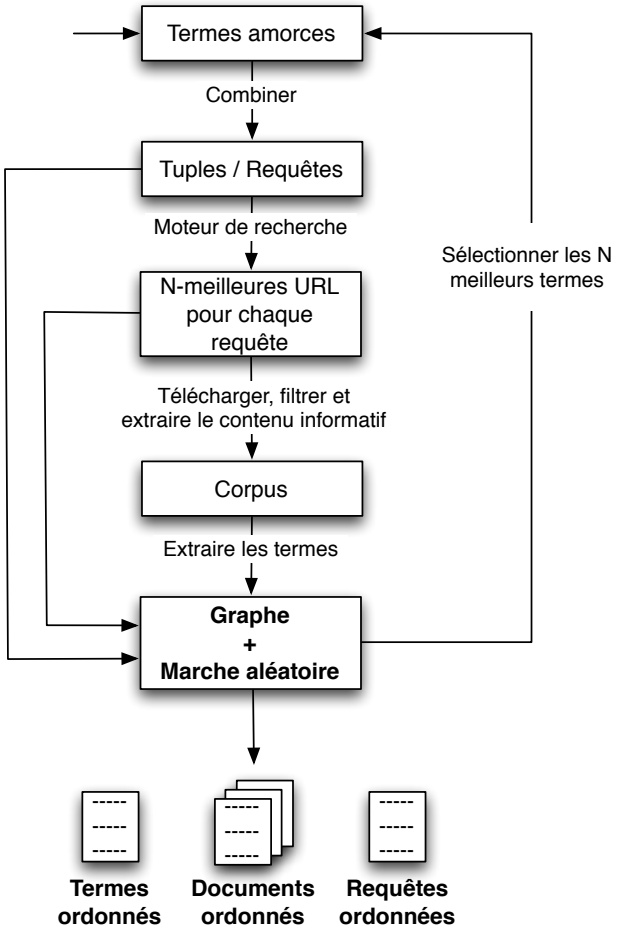

FIGURE 5.1: L'architecture du système proposé.

corpus ou de terminologies. En particulier, l'initiative *Web-as-Corpus WaCky* [18, 55, 140] a conduit à la création de très larges corpus web en plusieurs langues, grâce à la méthode BootCaT [17]. Cependant, la qualité des résultats étant fortement dépendante de la qualité des amorces, un filtrage manuel est nécessaire pour améliorer les performances. L'approche que nous proposons ici permet un filtrage automatiquement des termes s'écartant trop du domaine ou trop ambigus, ainsi que des documents trop éloignés, et apporte de meilleures performances pour la recherche. Le coût du filtrage manuel est ainsi très fortement réduit.

5.1 Graphes hétérogènes et bootstrapping

La figure 5.1 montre l'architecture de notre système. À partir de termes amorces fournies par l'utilisateur [1], des requêtes de n mots sont créées, fournies au moteur de recherche qui renvoie des documents, et de nouveaux termes sont extraits, qui serviront d'amorces pour une nouvelle itération.

Les termes, les documents et les requêtes sont conservés dans un graphe contenant l'ensemble des informations. Comme le montre la figure 5.2 (sur le thème *So-*

1. Ces amorces peuvent également être constituées de pages web, desquelles on extraiera une première liste de termes avec des techniques classiques d'extraction terminologique.

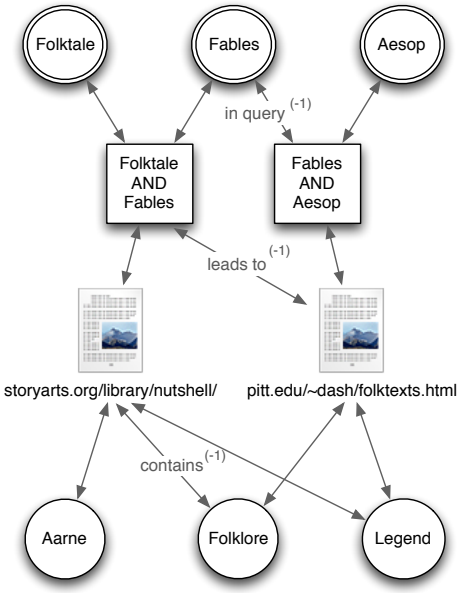

FIGURE 5.2: Une exemple de sous-graphe obtenu avec les amorces “boxoffice”, “Grammys” et “BBC”.

ciety/Folklore), les arcs de ce graphe peuvent représenter, selon les cas, les situations suivantes :

- Un *terme* entre dans la composition d'une *requête* ($in\_query(t, q)$) ;
- Une *requête* conduit à un *document* par l'intermédiaire du moteur de recherche généraliste ($leads\_to(q, d)$) ;
- Un *document* contient un *terme* ($contains(d, t)$). Ces arcs sont pondérés par une mesure de spécificité [82] telle que le *tf.idf* ou le *log odds ratio*. Nous normalisons alors systématiquement les poids des arcs pour nous assurer du caractère stochastique du graphe avant d'appliquer notre algorithme de marche aléatoire.

Comme proposé par [159], nous créons pour chaque relation r une relation inverse r^{-1}, ce qui modélise mieux l'influence réciproque existant entre les entités. De plus, cela garantit la convergence de l'algorithme de marche aléatoire employé vers une solution unique.

À la fin de chaque itération, cet algorithme de marche aléatoire, décrit à la section suivante, est appliqué, assignant un poids à chaque nœud du graphe, ce qui permet d'ordonner les documents et les termes [1]. La condition d'arrêt des itérations est généralement l'obtention d'un certain nombre de documents.

Notons par ailleurs que lorsque des pages web sont utilisées comme corpus, la précision et la robustesse des méthodes d'extraction employées dépendent fortement de la qualité du “nettoyage” de la page. Ce nettoyage est un domaine de recherche à part entière, et consiste notamment à extraire le contenu informatif d'une page (en retirant les scripts, les publicités et autres menus, dans notre cas avec l'outil BTE [57]), et à filtrer les doublons [28] ainsi que les pages d'erreur et de spam [58].

1. Les requêtes, étant elles-mêmes des nœuds, sont ordonnées également, même si ce n'est pas le résultat que nous souhaitons mettre en avant.

5.2 Marche aléatoire

Les algorithmes de marche aléatoire sont utilisés pour pondérer les nœuds d'un graphe en fonction des relations (arcs) qu'ils entretiennent avec les autres. Le poids d'un nœud sera d'autant plus important que les chemins qui mènent à lui seront nombreux et passeront par des nœuds eux-mêmes importants. On peut symboliser ce poids en imaginant un visiteur virtuel du graphe qui, placé à n'importe quel endroit de ce graphe, le parcourt en suivant les arcs au hasard. Certains nœuds, plus centraux, ont une probabilité plus élevée que d'autres d'être visités. C'est pourquoi on appelle cette classe d'algorithmes des marches aléatoires.

Ces méthodes ont été appliquées à de nombreux domaines du traitement automatique des langues, comme l'extraction de mots-clés ou de phrases [113], l'ordonnancement d'articles [122], de pages web [72, 124, 129], la désambiguïsation sémantique [2] ou la similarité sémantique [76]. Les algorithmes utilisés sont généralement des variantes plus ou moins éloignées du PageRank [124], utilisé pour mesurer l'importance d'une page web en utilisant les hyperliens entre les pages.

Si on considère un graphe orienté $G = (V, E)$, le score d'un nœud N_i est défini comme

$$PR(N_i) = (1 - \alpha)\lambda_0 + \alpha \times \sum_{j \in In(N_i)} \frac{PR(N_j)}{|Out(N_j)|}$$

où $In(N_i)$ (resp. $Out(N_i)$) sont les arcs entrants (resp. sortants) vers N_i, α un facteur d'amortissement et λ_0 un "vecteur de personnalisation", permettant d'orienter autoritairement le marcheur aléatoire vers certains nœuds préférentiels.

Dans l'algorithme original du PageRank [124], un facteur d'amortissement $\alpha = 0,85$ est utilisé, et le vecteur de personnalisation λ_0 est une distribution uniforme, c'est-à-dire qu'aucune préférence n'est indiquée. Nous conservons la valeur d'α, qui ne change d'ailleurs pas significativement les résultats finaux, mais nous suivons [129] et [72] en personnalisant notre marche pour que le marcheur reste à proximité des termes amorces :

$$\lambda_0 = (\sigma_0(t_0), \sigma_0(t_1), ..., \sigma_0(t_n))$$

$$\text{où } \sigma_0(t) = \begin{cases} 1.0 & \text{si } t \in \text{termes amorces} \\ 0.0 & \text{sinon} \end{cases}$$

Enfin, comme les arcs de notre graphe sont typés, nous modifions légèrement le choix des arcs par rapport à la version initiale (comme [159]) : au moment de choisir un nœud de destination (et donc un arc à suivre), l'utilisateur imaginaire choisit d'abord uniformément un type de relation, avant de choisir un arc correspondant à cette relation.

Après convergence de l'algorithme de marche aléatoire, nous obtenons pour chaque entité un poids reflétant l'importance globale de l'entité dans le graphe. Nous ordonnons alors chaque type d'entité indépendamment et obtenons une liste de documents, de termes et de requêtes ordonnés par importance relative. Cet ordre peut alors être utilisé pour filtrer les documents ou termes non pertinents plus facilement. À l'aide d'un seuil sur le nombre de documents ou sur le poids des documents, nous sélectionnons les nœuds-documents les plus centraux dans le graphe, c'est-à-dire les plus proches du thème représenté par les mots-clés amorces. Ces documents composeront la collection de référence de notre moteur de recherche thématique.

URL	http://www.cours-medecine.info/anatomie/
Titre	Cours d'Anatomie
Catégorie	`Santé/Médecine/Anatomie`
Description	Présente des cours d'anatomie humaine, des fiches de résumés, schémas et coupes à légender, ainsi que des QCM.

http://trec.nist.gov/
Text REtrieval Conference
`Computers/Software/Information_Retrieval`
An annual information retrieval conference and competition, the purpose of which is to support and further research within the information retrieval community.

http://www.geomagazine.fr/
GEO
`Loisirs/Voyage/Publications`
Magazine papier mensuel de reportages photographiques des quatre coins du monde. Sommaire et édito du numéro en kiosque, abonnement, index de tous les articles publiés.

http://www.cahiersducinema.com/
Les cahiers du cinéma
`Arts/Audiovisuel/Cinéma/Actualité_et_médias`
Version en ligne de la revue. Actualité, agenda, sorties et articles.

FIGURE 5.3: Exemple d'entrées de l'OpenDirectory.

5.3 Évaluation

Nous proposons (pour la première fois à notre connaissance) une évaluation objective et reproductible d'un outil de collecte de corpus thématique sur le Web, ne recourant pas à des moteurs de recherche dont nous ne maîtrisons ni l'index ni la fonction d'ordonnancement. Deux approches d'évaluation sont mises en œuvre, sur deux corpus différents : les pages web issues de l'OpenDirectory et des documents de l'Agence France Presse (AFP).

5.3.1 Évaluation sur l'OpenDirectory

L'OpenDirectory, parfois nommé DMOZ, est un répertoire de sites Web maintenu par une communauté d'éditeurs bénévoles, comportant plus de 5 millions d'entrées fin 2012, en 78 langues. Les sites sont organisés par thèmes hiérarchisés et décrits par quelques phrases de résumé. La figure 5.3 propose quelques exemples, et la figure 5.4 fournit un extrait de l'arborescence des catégories thématiques. Nous nous concentrons ici sur les catégories du second niveau.

Nous utilisons environ 2,3 millions de pages en anglais téléchargées à partir de ce répertoire. Ces pages sont ensuite indexées par le moteur Lucene [1]. Pour chaque catégorie, nous choisissons 10 mots-clés de façon automatique, en sélectionnant les termes ayant le *tf.idf* le plus élevé dans l'ensemble des textes de descriptions des sites. Des requêtes sont composées en combinant les mots-clés par paires ou par triplets.

1. http://lucene.apache.org

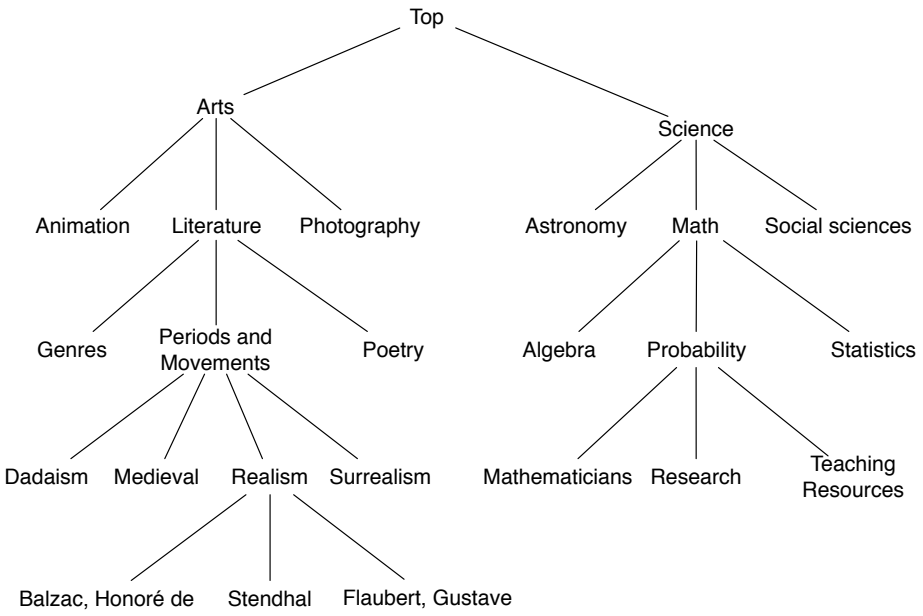

FIGURE 5.4: Extrait de l'arborescence de catégories de l'OpenDirectory.

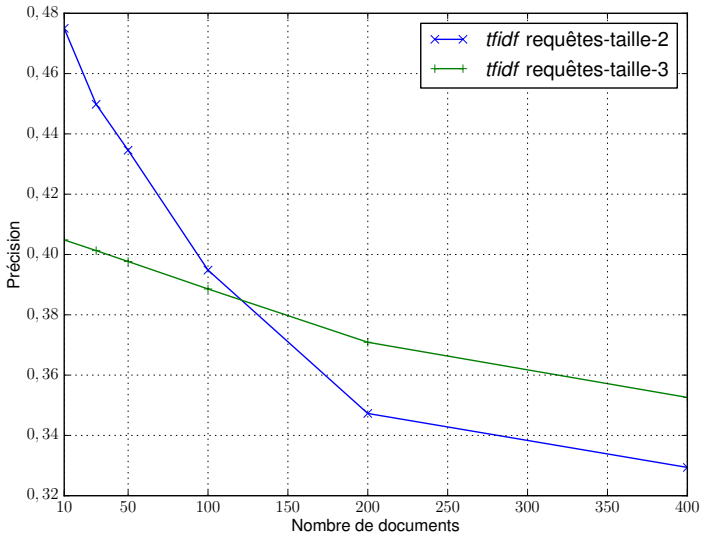

FIGURE 5.5: Évolution de la précision à k (nombre de documents retenus dans l'ordre suggéré par le système).

En utilisant ces termes en entrée du système, nous obtenons une liste ordonnée de documents et nous évaluant leur précision à k, avec k compris entre 10 et 400 documents retournés. Nous considérons qu'un document est pertinent s'il appartient dans l'OpenDirectory à la catégorie qui a servi à extraire les mots-clés. Nous présentons les résultats à la figure 5.5. La précision décroît au fur et à mesure de la sélection des documents dans la liste, ce qui semble indiquer que l'ordre des documents est valide.

Pour estimer l'apport de l'ordre produit, nous comparons également les performances obtenues avec l'ensemble des documents avec celles obtenues en ne conservant que les 100 documents les mieux classés. Nous voyons ainsi un gain en précision dans le second cas, de 0,065 pour les requêtes de taille 2 et de 0,036 pour les requêtes de

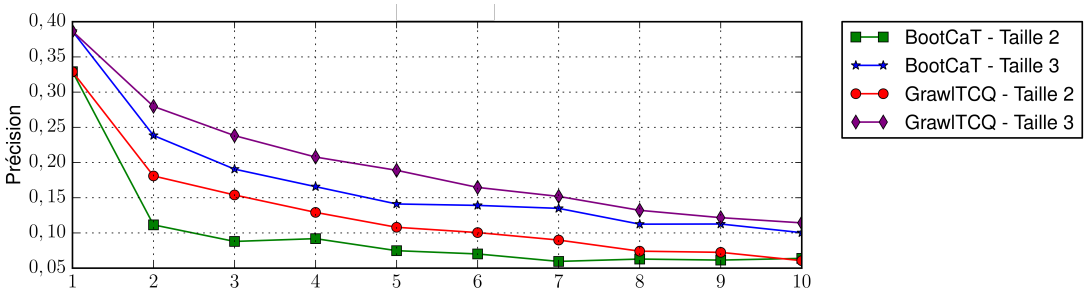

FIGURE 5.6: Évolution de la précision des corpus constitués au fur et à mesure des itérations.

Id	Category	#docs
01	Arts, culture and entertainment	3074
02	Crime, law and justice	5675
03	Disaster and accident	4602
04	Economy, business and finance	13321
08	Human interest	1300
11	Politics	17848
12	Religion and belief	1491
14	Social issue	1764
15	Sport	15089
16	Unrest, conflicts and war	8589

TABLE 5.1: Catégories du corpus AFP.

taille 3.

Nous étudions ensuite la robustesse de la sélection des mots-clés au fur et à mesure des itérations, en évaluant l'évolution de la précision après chaque nouvelle itération. En effet, chaque itération est supposée provoquer une dérive thématique plus ou moins importante, conduisant à une perte de précision. L'idée est qu'une bonne sélection de termes devrait faire baisser la précision de façon moins rapide qu'une mauvaise sélection.

Nous comparons notre approche avec la méthode standard BootCaT [17] sans filtrage manuel, ce qui constitue d'ailleurs la première évaluation de cette approche. À chaque itération, chacune des méthodes choisit deux nouveaux termes qui sont ajoutés à l'ensemble des amorces déjà choisies. Nous réalisons toutes les combinaisons de termes possibles de tailles 2 et 3 et nous téléchargeons les 10 meilleurs documents pour chaque requête. Nous évaluons la précision des documents récupérés par les deux méthodes au fur et à mesure des itérations. Notons que ce calcul ne tient pas compte de l'ordre produit par notre système. Par conséquent, dans cette expérience, seule la qualité des termes amorces à chaque itération influe sur la qualité des corpus créés. Sur les résultats présentés à la figure 5.6, nous constatons un gain obtenu par l'utilisation de la pondération des termes par la marche aléatoire.

Nous notons donc que la dérive thématique, inévitable dès lors que l'on utilise les techniques de *bootstrapping*, est réduite de manière significative, ce qui permet d'obtenir une diminution moins abrupte de la précision au fur et à mesure des itérations.

5.3.2 Évaluation sur le corpus de l'AFP

De manière complémentaire, nous utilisons également un corpus de l'AFP en anglais, composé de 57 441 dépêches écrites entre janvier et mars 2010. Ces documents

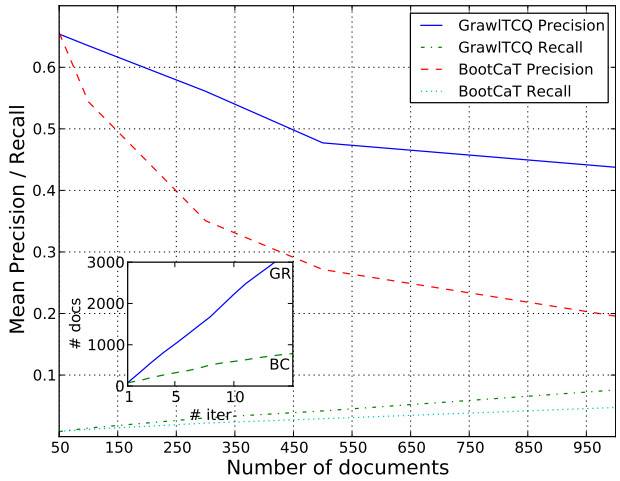

FIGURE 5.7: Précision moyenne et rappel à 50, 100, 300, 500 et 1000 documents (figure incise : nombre moyen de documents / nombre d'itérations).

sont annotés par leurs auteurs en mots-clés et en une ou plusieurs catégories IPTC [1], hiérarchie dont nous considérons ici le premier niveau (voir les 10 plus fréquentes à la table 5.1). Ceci nous permet d'évaluer l'approche dans un contexte thématique plus large, avec des mots-clés choisis par des experts et non automatiquement.

Le corpus a été indexé avec le moteur Lucene. De cette façon, la collection est stable et quantifiable ; les documents sont du contenu bien identifié, écrit dans un style journalistique ; ils sont déjà annotés par des personnes tierces, ce qui permet notamment d'effectuer autant d'évaluations automatiques que nécessaires, avec différents paramètres, contrairement à des résultats qui seraient jugés après coup.

Nous avons à nouveau utilisé l'algorithme BootCaT [17] comme *baseline*. Les deux algorithmes ont ainsi été comparé dans des conditions strictement identiques, par la tâche suivante : retrouver 50, 100, 300, 500 et 1000 documents pour chaque catégorie, indépendemment du nombre d'itérations.

Les termes amorces sont sélectionnés automatiquement grâce aux métadonnées des documents. La fréquence d'apparition d'un mot-clé dans une catégorie est divisée la somme de ses occurrences dans l'ensemble des catégories [3]. Les 10 mots ayant les valeurs les plus hautes sont conservés. Ainsi, on obtient des amorces pertinentes, qui ne sont pas exclusive à une catégorie. Par exemple, les termes sélectionnés pour la quatrième catégorie (économie et finance) sont (en anglais) *economics*, *summary*, *rate*, *opec*, *distress*, *recession*, *zain*, *jal*, *gold*, et *spyker*.

Chaque requête est composée de deux amorces, ce qui donne 45 combinaisons de requêtes, et un maximum de 10 documents par requête sont récupérés à chaque itération. Les résultats en termes de précision et de rappel selon le nombre de documents renvoyés sont présentés à la figure 5.7 et montrent les bonnes performances de notre système en comparaison de l'outil considéré comme l'état de l'art.

La description de ce travail de constitution automatique de corpus thématique achève notre partie dédiée au ciblage d'information. S'il peut sembler éloigné des autres parties, il est en revanche dans la continuité de nos activités de thèse de doctorat ; de plus, ce lien conservé avec ces problèmatiques ont façonné notre esprit d'une façon qui

1. http://www.iptc.org

contribue, nous l'espérons, à rendre notre projet de recherche original. Le chapitre qui suit, le dernier de ce mémoire, apporte quelques pistes quant à la poursuite de ce projet de recherche.

Chapitre 6

Conclusion et perspectives

Ce dernier chapitre est d'abord l'occasion de nous attarder sur quelques perspectives, quelques pistes que nous aimerions suivre autour du thème du traitement des événements et des informations temporelles, si l'avenir nous en accorde le temps et les moyens, et s'il ne nous réserve pas d'autres rencontres ou idées qui deviendront plus prioritaires. Nous le conclurons ensuite avec quelques réflexions plus générales, et parfois plus personnelles.

6.1 Perspectives

Le chapitre 1 a été l'occasion d'un tour d'horizon de la question des événements dans le traitement automatique des langues. En filigrane, s'y trouvait déjà quelques perspectives. Si nous reprenons le schéma récapitulatif présenté à la page 19, et reproduit ici dans le tableau 6.1, nous pouvons y voir une belle diagonale montrant la corrélation entre représentation de l'information et types de besoins d'information. En négatif, nous pouvons également faire le constat que la matrice contient beaucoup de trous. Sans tomber dans la surinterprétation d'un schéma à vocation principalement illustrative, nous pouvons tout de même voir dans ces trous autant de limites dans le traitement actuel de l'information.

En effet, la progression des modes de représentation, du "mot dans la phrase" jusqu'au "pic d'activité thématique", s'accompagne d'une perte de structure des événements, au fur et à mesure que la vision symbolique est remplacée par la vision statistique. Cette perte de structure est irréversible. Tout comme on ne peut recomposer un œuf une fois l'omelette préparée, il est difficile de revenir à une granularité fine de l'événement une fois qu'on l'a plongé dans un sac de mots. Pourtant, de nombreuses questions scientifiques et applications futures passeront de notre point de vue par une combinaison des approches, donnant au moins l'illusion à l'utilisateur qu'on recompose l'œuf pour lui, après en avoir fait une omelette. C'est en renforçant les liens entre d'une part le traitement automatique du langage, et, plus particulièrement, traitement automatique des descriptions textuelles d'événements, et d'autre part, des disciplines connexes comme la gestion des connaissances, les grandes masses de données, la fouille de données ou la visualisation d'information, que nous y parviendrons.

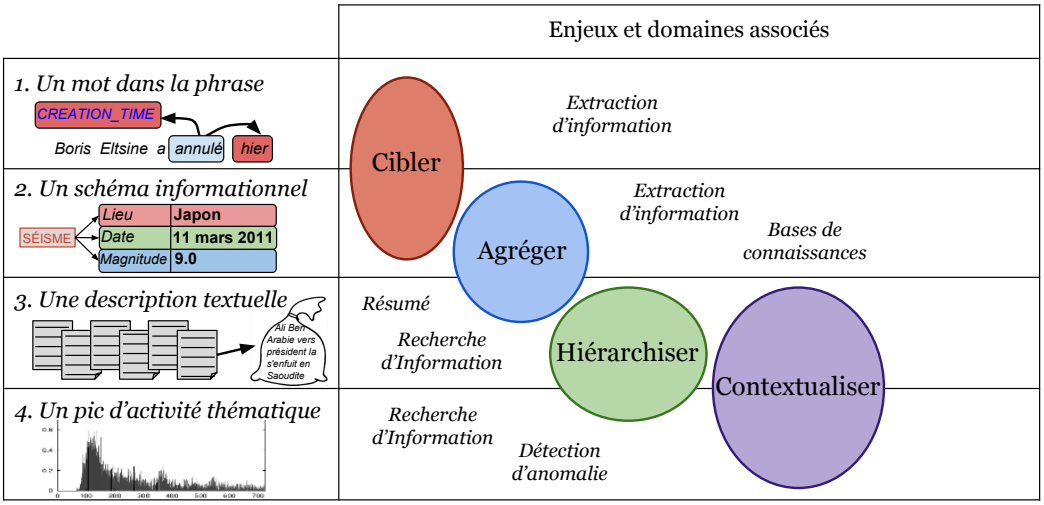

TABLE 6.1: Traitement des événements : enjeux et domaines associés (cf. Chapitre 1).

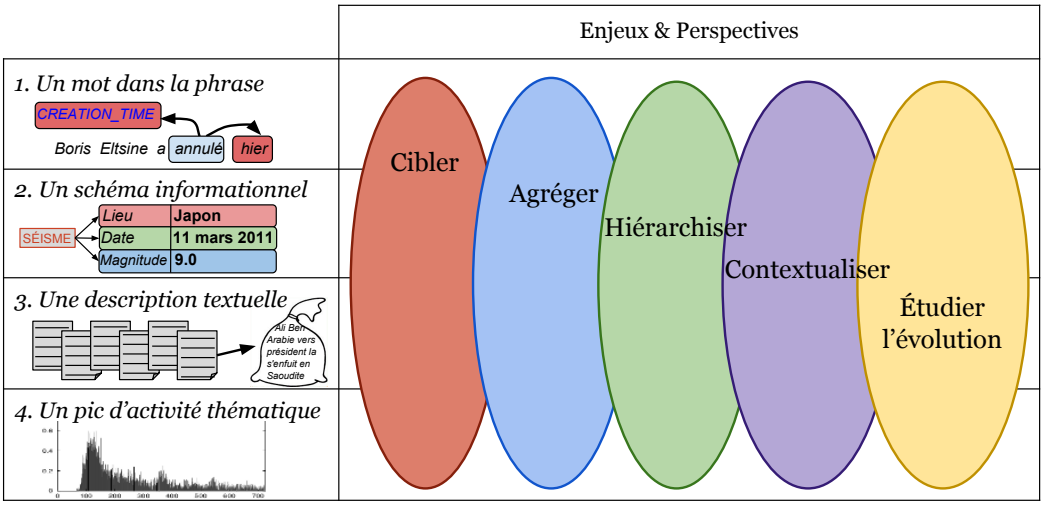

TABLE 6.2: Traitement des événements : perspectives.

6.1.1 Événements et bases de connaissances

Les bases de connaissances grossissent et se multiplient, pour le plus grand bénéfice de la science et du citoyen, notamment (dans le domaine général) grâce à l'ouverture massive des données institutionnelles et aux mouvements regroupés autour des termes *Open Data*, *Linked Data*, *Open Knowledge*. De son côté, la quantité d'information textuelle n'en est pas moins en forte augmentation, et la connaissance que l'on trouve de part et d'autre s'avère très complémentaire. Jamais les bases structurées ne pourront contenir l'ensemble de la connaissance, et l'accès à la sagesse (selon la fameuse pyramide DIKW [1]), par ailleurs fortement hypothétique, ne se fera en tout cas pas sans l'analyse du contenu textuel.

Les alliances entre les deux domaines sont d'ailleurs nombreuses [2], mais sont sans aucun doute amenées à prendre encore plus d'importance dans le futur. Dans le domaine journalistique, par exemple, où la question de la confiance est plus prégnante que jamais,

1. Voir la page 43.
2. Voir les campagnes KBP [79] mais aussi de nombreuses autres initiatives.

Fact-checking

la démarche récente du *fact-checking* (vérification des faits) tente d'assainir le débat démocratique, qui en a bien besoin. Chaque semaine nous amène son lot de polémiques, où les contre-vérités et les interprétations fallacieuses tiennent le haut du pavé. Les *fact-checkers* s'appuient encore sur des méthodes de recherche manuelles prenant beaucoup de temps et demandant une grande expertise de leur part. Pour analyser par exemple un discours politique et vérifier la véracité des assertions (données chiffrées, vérités assénées ou propos détournés), il est nécessaire de consulter des bases de données, de retrouver la source des citations, de comparer des interprétations différentes ou de remonter l'histoire d'un événement pour retrouver des informations plus objectives. Les méthodes existantes de traitement automatique des langues et de recherche d'information peuvent aider, par exemple pour analyser les citations, retracer le fil d'un événement, retrouver la source et distinguer la vérité de la croyance. De nouvelles approches doivent également être trouvées, par exemple pour apprendre du comportement du *fact-checker* face à une situation donnée, en fonction des textes à sa disposition et des bases de données qu'il consulte.

Dans un domaine spécialisé comme le domaine biomédical, auquel nous nous intéressons de plus en plus grâce au dynamisme de notre équipe en la matière, les liens entre les textes et les connaissances structurées sont également prépondérants. En effet, l'importance du domaine a conduit les chercheurs à construire de vastes terminologies permettant le lien entre des concepts tels que les gènes, les médicaments, les maladies, les examens, les opérations, les données anatomiques, etc. Mais les informations contenues dans les textes comme la littérature scientifique ou les dossiers patients restent très difficiles à apparier avec les concepts présents dans ces bases, et tout particulièrement les événements associés au parcours d'un patient. Pourtant, reconstituer automatiquement la chronologie des événements liés à un malade et être capable de faire le lien avec les connaissances cliniques en la matière constituerait une grande avancée pour l'aide au diagnostic et aux décisions de traitement.

Parcours de soin

6.1.2 Événements et masses de données

Ce dernier point est également en lien avec la question récurrente de la quantité et de la redondance de l'information. Pour rester dans le domaine des chronologies du domaine médical, la généralisation récente des dossiers électroniques patients (comptes-rendus d'hospitalisation, ordonnances, courriers entre praticiens...) apporte à la fois la redondance — les événements marquants y sont répétés de nombreuses fois — et la quantité — des milliers de dossiers sont disponibles dans les hôpitaux. Comme pour le domaine journalistique, beaucoup abordé dans ce mémoire, la redondance devra permettre de hiérarchiser l'information et d'affiner la chronologie d'un parcours de soin, en donnant plus d'importance aux événements fréquemment mentionnés et en découvrant des liens entre événements. On pourra ainsi construire progressivement, au fur et à mesure de l'arrivée de nouveaux documents, un graphe temporel pondéré de la prise en charge. Au-delà de cela, la quantité en termes de nombre de dossiers ouvre la voie, à plus long terme, à des études de fouille des données extraites, pour découvrir des liens insoupçonnés entre symptômes, traitements, opérations et autres actions.

Fouille de données

De façon plus générale, l'aspect multidocument de l'analyse des événements a déjà été largement traité dans les chapitres précédents, et nous pensons avoir ouvert quelques voies dans les travaux décrits au chapitre 3 qui méritent d'être creusées plus en profondeur. En particulier, certaines notions nécessitent un effort de formalisation. Tout comme la vision "noyau/argument" typique de l'extraction d'information est relativement claire, tout comme les événements et les expressions temporelles au niveau du texte ont donné lieu à un travail important sur le langage de représentation TimeML, une démarche équivalente permettrait d'apporter un cadre à la vision que nous avons

Formalisme appelée “événement dans le monde”. Dans cette optique, un événement n’est plus seulement “quelque chose qui se passe”, mais aussi quelque chose dont on parle, qui possède un contexte, des causes, un déroulement, des conséquences, dont la désignation évolue et dont les connaissances que l’on a à son sujet évoluent également. Nous ne prétendons pas qu’aucune réflexion n’a été menée à ce sujet, et d’ailleurs les mots que nous venons d’énumérer sont ou pourraient être des noms de relations dans des théories linguistiques bien connues ; mais le cadre n’est plus du tout le même lorsqu’on s’attaque à la complexité et à la multitude des sources et des types de documents.

Évolution Nous avons également argumenté au chapitre 1 pour une étude plus poussée des aspects concernant l’évolution des connaissances concernant les événements. Évolution d’une situation médicale, évolution d’un événement médiatique, de son importance, des informations le concernant, de ses conséquences, de sa désignation linguistique. Ici encore, des collaborations avec des spécialistes de fouille de données dynamiques seront indispensables, pour trouver des angles d’étude à la fois pertinents et raisonnables.

Induction de schémas Enfin, la quantité d’information textuelle peut également permettre d’espérer s’affranchir de la limite traditionnelle de l’extraction d’information, pour laquelle une définition *a priori* des schémas d’extraction (liste d’attributs) est nécessaire. Dans le cadre des pratiques naissantes de l’*Open Information Extraction* et de l’induction de schémas [15, 92, 34, 32], l’objectif est d’apprendre à découvrir les attributs pertinents pour un type d’événement donné. Dans cette optique, à partir de mots-clés amorces représentant le thème (exemple : *séisme*), le système déduit les caractéristiques importantes (*magnitude*, *épicentre*, etc.). Celles-ci sont ensuite utilisées pour constituer une base rassemblant les instances des événements ciblés. Un projet dans ce domaine a débuté très récemment sous notre coordination.

6.1.3 Événements et visualisation

Journalisme de données Pour terminer, nous croyons également beaucoup à l’alliance du traitement automatique des langues et du journalisme de données, après les travaux préliminaires mais très encourageants présentés à la section 3.3. Là encore, une réflexion sur la formalisation et la systématisation des approches semble indispensable, et des collaborations plus poussées avec des journalistes et des spécialistes de la visualisation doivent être mises en place. Les questions qui se posent sont notamment :

- Quels types d’information peuvent être recueillis avec des méthodes comparables à celle présentée ? Ici les limites sont celles de la complexité linguistique et de l’imagination.
- Quelles sont les approches de TAL compatibles avec l’extraction de données statistiques à partir des textes ? La vision privilégiant la précision d’abord, que nous avons adopté dans notre travail, n’est pas la seule possible, et une perspective de rappel d’abord suivi d’un regroupement ou d’un filtrage par apprentissage non supervisé, par exemple, semble pouvoir convenir aussi et s’adapter plus facilement à de nombreux problèmes.
- Quels sont les traitements pertinents des données obtenues par l’extraction d’information ? Dans notre première étude, nous avons vu que les sorties brutes conduisaient à des informations trop bruitées, mais qu’un lissage adéquat suffisait à les rendre beaucoup plus significatives.
- Quelles sont les applications envisageables pour des domaines de spécialité ?
- Quels sont les liens potentiels avec la question du *fact-checking* abordée plus haut ?

6.2 Conclusion

L'ensemble des travaux décrits ici, ainsi que les perspectives, visent à répondre à certains types de besoins d'informations. Ils s'inscrivent dans une frange particulière du spectre scientifique, que certains appellent "recherche appliquée". D'un côté, le grand public, matérialisé par nos familles et amis (*"Alors, tu trouves ?"*), se pose de façon récurrente et légitime la question de la finalité des recherches qu'il finance ("*À quoi ça sert ?*"), et constate que ses besoins d'information quotidiens sont couverts par des multinationales américaines ; il considère souvent que notre but et notre rôle sont de renverser cette situation, et voit bien que ceci n'est pas proche d'arriver ("*De toute façon, Google le fera mieux que vous, alors à quoi bon ?*"). De l'autre côté, certains scientifiques estiment que notre tâche principale de chercheur est d'enrichir la connaissance et que, sans mépris particulier pour la valorisation, la recherche d'applications n'est pas de notre ressort (là aussi, le débat est récurrent, du coin café aux commentaires des relecteurs de nos articles, en passant par notre instance d'évaluation préférée).

À ces derniers, nous répondons que si nos réflexions ne consistaient qu'à prendre des méthodes et des idées décrites par d'autres pour les appliquer à un problème spécifique, sans apporter d'autre plus-value que la satisfaction d'un groupe d'utilisateurs concerné, alors nous ne ferions en effet pas ce que l'on attend d'un chercheur du service public. En revanche, en proposant des nouvelles méthodes pour traiter des problèmes, ou en ouvrant de nouvelles portes et en inventant de nouvelles tâches, puis en montrant avec une démarche scientifiquement solide et une évaluation significative que ces idées permettent de mieux répondre aux besoins d'information du citoyen ou du spécialiste, il nous semble que nous sommes pleinement dans notre rôle. Cette question de l'évaluation est d'ailleurs centrale, nous devons la garder à l'esprit en permanence, voire être innovant en la matière, en proposant des nouvelles méthodologies et en organisant des campagnes d'évaluation.

Devant nos familles et amis, nous admettons qu'en effet l'histoire retient le nom de ceux qui ont permis à une invention d'entrer dans les foyers et de changer la vie des gens. Nous leur expliquons surtout que de l'imprimerie au Web, toutes les grandes avancées ont été le fruit d'un bouillonnement scientifique et intellectuel, auquel des centaines de personnes ont apporté une pierre plus ou moins modeste, et qui à la fin profite pour la postérité aux plus opportunistes, aux plus audacieux ou aux plus chanceux. Nous concluons que malgré l'importance de l'ego dans la recherche, cette position d'ingrédient chimique contribuant à la catalyse nous convient, et que nous sommes encore dans un domaine dans lequel avec des moyens raisonnables, une bonne idée peut faire avancer les choses, même si l'on n'a ni accès à 1000 milliards de pages web, ni besoin de créer une société fantôme en Irlande pour payer moins d'un milliard d'euros d'impôts.

Enfin, je leur réponds également que j'enseigne. Cela résoud d'emblée à leurs yeux la question de la place dans la société (même si, en contrepartie, ils en concluent immédiatement que j'ai cinq mois de vacances par an). Aux plus motivés, j'explique que cela ne fait pas de moi un chercheur à mi-temps, mais un enseignant-chercheur. Qu'à titre personnel, je trouve dans l'enseignement l'équilibre et le lien avec la vie réelle, ainsi qu'une équipe tournée vers un but commun (faire tourner la boutique avec les moyens du bord) ; tandis que la recherche m'apporte l'excitation nécessaire à la mission, s'occupe de mon ego (avec des hauts et des bas, bien entendu) et m'offre des collaborations passionnantes guidées par des affinités et un intérêt communs. Qu'à titre professionnel, même si le quotidien s'apparente à une recherche permanente de compromis entre les activités de recherche et d'enseignement, il s'avère quand on y réfléchit que les liens entre elles sont bien plus nombreux qu'on pourrait le penser, bien au-delà des quelques heures annuelles de cours de Master 2.

Ce mémoire a été l'occasion de présenter des recherches menées en extraction d'information et en fouille de textes des dernières années. Certains travaux plus anciens ou plus à la marge ont été passés sous silence, pour maintenir une certaine unité thématique.

Toutes les approches proposées s'inscrivent dans au moins un des quatre axes de recherche identifiés, dans la quête de réponse aux divers besoins d'information, à savoir le ciblage, l'agrégation, la hiérarchisation et la contextualisation d'information [1]. Certaines tentent de les combiner ou de les fusionner, ce qui me semble constituer une vision intéressante pour l'avenir. J'ai enfin présenté quelques pistes supplémentaires, dans la continuité de mes travaux récents. S'il est impossible, à la vitesse à laquelle évoluent la société et la recherche, de prétendre que ces pistes vont occuper une carrière entière, elles devraient tout de même être au centre de mes préoccupations dans les années à venir.

1. Sur ce dernier point, voir également la participation dans l'organisation de la tâche "Tweet Contextualization" à INEX, de 2010 à 2013 [22, 135, 81, 80].

Bibliographie

[1] E. Agichtein and L. Gravano. Snowball : Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, 2000. 12

[2] E. Agirre and A. Soroa. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL, 2009*, 2009. 88

[3] K. Ahmad, L. Gillam, and L. Tostevin. University of surrey participation in TREC 8 : Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILTER). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 1999. 92

[4] S. Aït-Mokhtar, J.-P. Chanod, and C. Roux. Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering*, 8 :121–144, 2002. 26, 48, 77

[5] J. Allan, editor. *Topic Detection and Tracking*. Springer, 2002. 17, 46, 53, 71

[6] J. Allen. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, pages 832–843, 1983. 8, 21, 35

[7] J. Allen, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 10–18, 2001. 46

[8] B. Arnulphy. *Désignations nominales des événements : Étude et extraction automatique dans les textes*. PhD thesis, Université Paris-Sud, OCT 2012. 23

[9] B. Arnulphy, V. Claveau, X. Tannier, and A. Vilnat. Techniques d'apprentissage supervisé pour l'extraction d'événements TimeML en anglais et français. In *Actes de la COnférence en Recherche d'Information et Application (CORIA 2014)*, Nancy, France, 2014. 30

[10] B. Arnulphy and X. Tannier. Entités Nommées Événement : guide d'annotation. Technical Report 2013-12, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), jul 2013. 24

[11] B. Arnulphy, X. Tannier, and A. Vilnat. Les entités nommées événement et les verbes de cause-conséquence. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2010, article court)*, Montréal, Canada, July 2010. 22, 27

[12] B. Arnulphy, X. Tannier, and A. Vilnat. Un lexique pondéré des noms d'événements en français. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2011, article court)*, Montpellier, France, July 2011. 22

[13] B. Arnulphy, X. Tannier, and A. Vilnat. Automatically Generated Noun Lexicons for Event Extraction. In *Proceedings of CicLing 2012*, mar 2012. 23, 26

[14] B. Arnulphy, X. Tannier, and A. Vilnat. Event Nominals : Annotation Guidelines and a Manually Annotated Corpus in French. In ZLREC2012 [167]. 24

[15] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2670–2676, Hyderabad, India., 2007. 12, 98

[16] J. Barnes and P. Hut. A hierarchical O(N log N) force-calculation algorithm. *Nature*, 324 :446–449, dec 1986. 67

[17] M. Baroni and S. Bernardini. BootCaT : Bootstrapping Corpora and Terms from the Web. In (ELRA) [53], pages 1313–1316. 86, 91, 92

[18] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3) :209–226, 2009. 86

[19] R. Barzilay, K. R. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA, 1999. Association for Computational Linguistics, Morristown, NJ, USA. 16

[20] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics : Real-World Event Identification on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011. 17

[21] C. Bejan and S. Harabagiu. Unsupervised Event Coreference Resolution. *Computational Linguistics*, 2013. 30

[22] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, and X. Tannier. Overview of the INEX Tweet Contextualization 2013 Track. In *Working Notes for the CLEF 2013 Workshop*, Valencia (Spain), Sept. 2013. 100

[23] S. Bethard and J. H. Martin. Identification of Event Mentions and their Semantic Class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 146–154, Sydney, Australia, 2006. Association for Computational Linguistics. 22

[24] A. Bittar. *Building a TimeBank for French : A Reference Corpus Annotated According to the ISO-TimeML Standard*. PhD thesis, Université Paris Diderot, 2010. 24

[25] E. Boschee, R. Weischedel, and A. Zamanian. Automatic information extraction. In *Proceedings of the International Conference on Intelligence Analysis*, 2005. 71

[26] G. Bouma, I. Fahmi, J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann. Linguistic knowledge and question answering. *Traitement Automatique des Langues*, 46(3) :15–39, 2007. 78

[27] T. Brants, F. Chen, and A. Farahat. A System for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 330–337, Toronto, Canada, 2003. ACM. 46

[28] A. Z. Broder. Identifying and Filtering Near-Duplicate Documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, pages 1–10, London, UK, 2000. Springer-Verlag. 87

[29] C. Bron and J. Kerbosch. Algorithm 457 : finding all cliques of an undirected graph. *Communications of the ACM*, 16(9) :575–577, 1973. 58

[30] L. Calabrese. Les héméronymes. Ces événements qui font date, ces dates qui deviennentévénements. *Mots. Les langages du politique*, 3 :115–128, 2008. 23, 28, 30

[31] L. Calabrese. La nomination d'événements dans le discours d'information : entre activité collective et déférence épistémiologique. In *Colloque Langage, discours, événements*, Firenze, Italy, 2011. 22, 30

[32] N. Chambers. Event Schema Induction with a Probabilistic Entity-Driven Model. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP 2013)*, Seattle, USA, 2013. 98

[33] N. Chambers and D. Jurafsky. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08 : HLT*, pages 789–797, Columbus, Ohio, June 2008. Association for Computational Linguistics, Morristown, NJ, USA. 12, 13

[34] N. Chambers and D. Jurafsky. Template-Based Information Extraction without the Templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, page 976–986, Portland, USA, 2011. 98

[35] A. X. Chang and C. Manning. SUTime : A library for recognizing and normalizing time expressions. In ZLREC2012 [167]. 6

[36] M. Charolles. *La référence et les expressions référentielles en français*. Collection L'essentiel Français. Ophrys, Paris, France, 2002. 5

[37] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 425–432, Sheffield, United Kingdom, 2004. 46

[38] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2005)*, Edinburgh, UK, 2005. 13

[39] F. Cicurel. Pré-visibilité des Discours Journalistiques : À Propos d'un Événement-Catastrophe. *Les Carnets du Cediscor*, 2009. 15

[40] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 43(6) :551–558, 1960. 56

[41] S. Cohen, J. T. Hamilton, and F. Turner. Computational Journalism. *Communications of the ACM*, 54(11) :66–71, 2011. 60

[42] L. Danlos and B. Gaiffe. Event coreference and discourse relations. In L. Kulda, editor, *Language, Music and Cognition*. Kluwer Academic Publisher, 2004. 30

[43] D. Davidson. *Essays on Actions and Events*. Psychology as Philosophy. Calendron Press, Oxford, UK, 1980. Chapitre 11 "Mental Events" (1970). 5

[44] D. Davidson. *Actions et Événements*. Épithémée. Presses Universitaires de France, Paris, France, Dec. 1993. (traduit de l'américain par Pascal Engel). 5

[45] D. Day, R. Kozierok, C. McHenry, and L. D. Riek. Callisto : A Configurable Annotation Workbench. In (ELRA) [53]. 28

[46] C. de Groc and X. Tannier. Experiments on Pseudo Relevance Feedback using Graph Random Walks. In *Proceedings of the 19th International Symposium on String Processing and Information Retrieval (SPIRE 2012, Short Paper)*, volume LNCS 7608, pages 193–198, Cartagena, Colombia, Oct. 2012. 85

[47] C. de Groc and X. Tannier. Apprendre à ordonner la frontière de crawl pour le crawling orienté. In *Actes de la COnférence en Recherche d'Information et ses Applications (CORIA 2014)*, Nancy, France, Mar. 2014. 85

[48] C. de Groc, X. Tannier, and J. Couto. GrawlTCQ : Terminology and Corpora Building by Ranking Simultaneously Terms, Queries and Documents using Graph Random Walks. In *Proceedings of ACL Workshop on Graph-based Methods for Natural Language Processing (TextGraph 2011)*, pages 37–41, Portland, Oregon, USA, July 2011. 85

[49] C. de Groc, X. Tannier, and C. de Loupy. Un critère de cohésion thématique fondé sur un graphe de cooccurrences. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2012)*, pages 183–195, Grenoble, France, June 2012. 85

[50] P. Denis and P. Muller. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1788–1793, Barcelona, Spain, 2011. 6

[51] N. Diakopoulos. Cultivating the Landscape of Innovation in Computational Journalism. CUNY Tow-Knight Center for Entrepreneurial Journalism, apr 2012. 60

[52] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web Question Answering : Is More Always Better ? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 291–298, Tampere, Finland, 2002. ACM. 70

[53] E. L. R. A. (ELRA), editor. *Proceedings of the Fourth International Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. 102, 104

[54] A. Feng and J. Allan. Incident Threading for News Passages. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China, nov 2009. 53

[55] A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*, pages 47–54, 2008. 86

[56] K. Filippova. Multi-sentence compression : finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 322–330, Beijing, China, Aug. 2010. 16

[57] A. Finn, N. Kushmerick, and B. Smyth. Fact or fiction : Content classification for digital libraries. In *Proceedings of the joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries*, 2001. 87

[58] W. H. Fletcher. Making the Web More Useful as a Source for Linguistic Corpora. In *Corpus Linguistics in North America*, pages 191–205, 2004. 87

[59] J. Fragnon. Quand le 11-Septembre s'approprie le onze septembre. Entre dérive métonymique et antonomase. *Mots. Les langages du politique.*, 85 :82–95, 2007. 30

[60] M. Freedman, L. Ramshaw, E. Boschee, R. Gabbard, G. Kratkiewicz, N. Ward, and R. Weischedel. Extreme Extraction – Machine Reading in a Week. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2011)*, pages 1437–1446, Edinburgh, UK, 2011. 12

[61] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1) :119–139, 1997. 51

[62] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *VLDB '05 : Proceedings of the 31st international conference on Very large data bases*, pages 181–192, Trondheim, Norway, 2005. 18, 46

[63] L. Galescu and N. Blaylock. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 19th annual International Forum on Quality and Safety in Healthcare*, pages 715–720, Paris, France, apr 2012. 9

[64] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 40–48, Seattle, Washington, 2000. Association for Computational Linguistics. 15

[65] R. Grishman and B. Sundheim. Message Understanding Conference - 6 : A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 466–471, Copenhagen, Danmark, 1996. 10

[66] C. Grouin, N. Grabar, T. Hamon, S. Rosset, X. Tannier, and P. Zweigenbaum. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association*, Apr. 2013. 35

[67] Y. Gu, J. Cui, H. Liu, X. Jiang, J. He, X. Du, and Z. Li. Detecting hot events from web search logs. In *Proceedings of the 11th international conference on Web-age information management (WAIM'10)*, pages 417–428, Jiuzhaigou, China, 2010. Springer-Verlag. 17

[68] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. W. 1. The WEKA Data Mining Software : An Update. *SIGKDD Explorations*, 11(1), 2009. 57, 64

[69] H. Harkema, A. Setzer, R. Gaizauskas, and M. Hepple. Mining and Modelling Temporal Clinical Data. In *The UK e-Science All Hands Meeting Proc*, pages 507–514, 2005. 8

[70] S. Hartrumpf, I. Glückner, and J. Leveling. University of Hagen at QA@CLEF 2008 : Efficient Question Answering with Question Decomposition and Multiple Answer Streams. In ZCLEF08 [164]. 76

[71] E. Hatcher and O. Gospodnetić. *Lucene in Action*. Manning, 2004. 80

[72] T. H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the eleventh international conference on World Wide Web (WWW'02)*, page 517, New York, New York, USA, 2002. ACM Press. 88

[73] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational linguistics (COLING'92)*, pages 539–545, Nantes, France, 1992. 12

[74] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2 : A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence, special issue on Wikipedia and Semi-Structured Resources*, 2013. 60

[75] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *Proceedings of KDD*, page 168–177, Seattle, USA, 2004. 71

[76] T. Hughes and D. Ramage. Lexical Semantic Relatedness with Random Graph Walks. In *Proceedings of EMNLP, 2007*, pages 581–589, 2007. 88

[77] C. Jacquemin. What is the tree that we see through the window : A linguistic approach to windowing and term variation. *Information Processing & Management*, 32(4) :445–458, 1996. 78

[78] L. Jean-Louis, R. Besançon, and O. Ferret. Using Temporal Cues for Segmenting Texts into Events. In *Proceedings of IceTAL 2010*, pages 150–161, 2010. 15

[79] H. Ji and R. Grishman. Knowledge Base Population : Successful Approaches and Challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 1148–1158, Portland, Oregon, 2011. Association for Computational Linguistics. 96

[80] E. S. Juan, P. Bellot, V. Moriceau, and X. Tannier. Overview of the INEX 2010 Question Answering Track (QA@INEX). In S. Geva, J. Kamps, R. Schenkel, and A. Trotman, editors, *Comparative Evaluation of Focused Retrieval, 9th International Workshop of the Inititative for the Evaluation of XML Retrieval, INEX 2010*, volume LNCS 6932 of *Lecture Notes in Computer Science*, pages 269–281, Vugh, The Netherlands, 2011. 100

[81] E. S. Juan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2011 Question Answering Track (QA@INEX). In S. Geva, J. Kamps, and R. Schenkel, editors, *Focused Retrieval of Content and Structure, 10th International Workshop of the Inititative for the Evaluation of XML Retrieval, INEX 2011*, volume LNCS 7424 of *Lecture Notes in Computer Science*, pages 269–281, 2012. 100

[82] K. Kageura and B. Umino. Methods of automatic term recognition : A review. *Terminology*, 3(2) :259–289, 1996. 87

[83] N. Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelone, Spain, 2004. Association for Computational Linguistics, Morristown, NJ, USA. 71

[84] B. Katz, G. Borchardt, and S. Felshin. Syntactic and Semantic decomposition Strategies for Question Answering from Multiple Resources. In *Proceedings of the AAAI 2005 Workshop on Inference for Textual Question Answering*, Pittsburgh, Pennsylvania, USA, 2005. 76

[85] B. Katz and J. Lin. Selectively Using Relations to Improve Precision in Question Answering. In *Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering*, Apr. 2003. 76

[86] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3) :637–649, 2001. 57, 64

[87] R. Kessler, X. Tannier, C. Hagège, V. Moriceau, and A. Bittar. Finding Salient Dates for Building Thematic Timelines. In *50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 730–739, Jeju Island, Korea, 2012. 14, 16, 45, 46

[88] R. Kessler, X. Tannier, C. Hagège, V. Moriceau, and A. Bittar. Extraction de dates saillantes pour la construction de chronologies thématiques. *Traitement Automatique des Langues, numéro spécial sur le traitement automatique des informations temporelles et spatiales*, 53(2), 2013. 22, 45, 46

[89] T. Kitani, Y. Eriguchi, and M. Hara. Pattern Matching and Discourse Processing in Information Extraction from Japanese Text. *Journal of Artificial Intelligence Research*, 2 :89–110, 1994. 15

[90] D. Klein and C. D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 423–430, Sapporo, Japan, jul 2003. Association for Computational Linguistics. 63

[91] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate Detection using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM)*, New York, USA, 2010. 62

[92] S. Kok and P. Domingos. Extracting Semantic Networks from Text Via Relational Clustering. In *Proceedings of ECML PKDD'08*, pages 624–639, 2008. 12, 98

[93] R. Kuć. *Apache Solr 4 Cookbook*. Packt, 2013. 61, 66

[94] G. Kumaran and J. Allen. Text classification and named entities for new event detection. In *SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304. ACM, 2004. 46

[95] A. Lascarides and N. Asher. Segmented Discourse Representation Theory : Dynamic Semantics with Discourse Structure. In H. Bunt and R. Muskens, editors, *Computing Meaning : Volume 3*, pages 87–124. Kluwer Academic Publishers, 2007. 21

[96] M. Lecolle. Éléments pour la caractérisation des toponymes en emploi événementiel. In I. Evrard, M. Pierrard, L. Rosier, and D. V. Raemdonck, editors, *Les sens en marge - Représentations linguistiques et observables discursifs : actes du colloque international de Bruxelles, 3-5 novembre 2005*, pages 29–43. L'Harmattan, 2009. 5, 23, 30

[97] H. Lee, M. Recasens, A. Chang, M. Surdeanu, and D. Jurafsky. Joint Entity and Event Coreference Resolution across Documents. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP 2012)*, Jeju, South Korea, 2012. 30

[98] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2013. 12, 60, 62

[99] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A Probabilistic Model for Restrospective News Event Detection. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005. ACM Press, New York City, NY, USA. 46

[100] A.-L. Ligozat, B. Grau, I. Robba, and A. Vilnat. Systèmes de questions-réponses : vers la validation automatique des réponses. In *14ème Traitement automatique des langues naturelles*, pages 173–182, Toulouse, France, June 2007. ATALA. 76

[101] C.-Y. Lin and E. Hovy. From single to multi-document summarization : A prototype system and its evaluation. In ZACL02 [163], page 457–464. 46

[102] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. 71

[103] H. Llorens, L. Derczynski, R. Gaizauskas, and E. Saquete. TIMEN : An Open Temporal Expression Normalisation Resource. In ZLREC2012 [167]. 6

[104] N. Lucas. La rhétorique des dépêches de presse à travers les marques énonciatives du temps, du lieu et de la personne. In *Actes de la Semaine du Document Numérique (SDN 2004), Journée ATALA.*, 2004. 15

[105] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, P. Osenova, A. Peñas, V. Jijkoun, B. Sacaleanu, P. Rocha, and R. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answering Track. In *Working Notes QA@CLEF*, Alicante, 2006. 82

[106] B. Magnini, M. Negri, and H. Tanev. Is It the Right Answer ? Exploiting Web Redundancy for Answer Validation. In ZACL02 [163], page 425–432. 70

[107] I. Mani and B. Schiffman. Temporally Anchoring and Ordering Events in News. In J. Pustejovsky and R. Gaizauskas, editors, *Time end Event Recognition in Natural Language*. John Benjamin, 2005. 6

[108] W. C. Mann and S. A. Thompson, editors. *Discourse Description. Diverse linguistic analyses of a fund-raising text*. John Benjamins, 1992. 21

[109] M.Banko, E.Brill, S.Dumais, and J.Lin. AskMSR : Question Answering Using the Worldwide Web. In *Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, page 7–9, Palo Alto, USA, 2002. 70

[110] K. McKeown, R. Barzilay, S. Blaire-Goldensohn, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman. The Columbia Multi-Document Summarizer for Document Understanding Conference. In *Proceedings of the Document Understanding Conference (DUC 2002)*, 2002. 15

[111] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the second international conference on Human Language Technology Research (HLT'02)*, pages 280–285, San Diego, California, USA, 2002. Morgan Kaufmann Publishers Inc. 14

[112] K. R. McKeown and D. R. Radev. Generating summaries of multiple news articles. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 74–82, Seattle, Washington, USA, July 1995. ACM Press, New York City, NY, USA. 14, 46

[113] R. Mihalcea and P. Tarau. TextRank bringing order into text. In *Proceedings of EMNLP*, pages 404–411. Barcelona : ACL, 2004. 88

[114] G. A. Miller. WordNet : a lexical database for English. *Communications of the ACM*, 38(11) :39–41, Nov. 1995. 26

[115] S. Miller, H. Fox, L. Ramshaw, and R. Weischedel. A novel use of statistical parsing to extract information from text. In *Proceedings of the Conference of the North-American Association for Computational Lingguistics (NAACL 2000)*, 2000. 71

[116] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2 - Volume 2*, pages 1003–1011, Suntec, Singapore, 2009. Association for Computational Linguistics. 12

[117] S. Moirand. *Les discours de la presse quotidienne. Observer, analyser, comprendre*. PUF, 2007. Isbn : 2-13-055923-9. 5

[118] V. Moriceau and X. Tannier. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In *Proceedings of the 9th International Language Resources and Evaluation (LREC'2014)*, Reykjavik, Iceland, May 2014. 48

[119] V. Moriceau, X. Tannier, A. Grappy, and B. Grau. Justification of Answers by Verification of Dependency Relations - The French AVE Task. In ZCLEF08 [164]. 81

[120] P. Muller and X. Tannier. Annotating and measuring temporal relations in texts. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling 04)*, pages 50–56, Geneva, Switzerland, Aug. 2004. COLING. 35

[121] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event Threading within News Topics. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2004)*, Washington DC, USA, nov 2004. 53

[122] Z. Nie, Y. Zhang, J. Wen, and W. Ma. Object-level ranking : Bringing order to web objects. In *Proceedings of the 14th international conference on World Wide Web*, pages 567–574. ACM, 2005. 88

[123] C. Nédellec, R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria, aug 2013. Association for Computational Linguistics. 13

[124] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking : Bringing Order to the Web. Technical report, Stanford InfoLab, 1999. 88

[125] J. C. Platt. *Advances in Kernel Methods - Support Vector Learning*, chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization. MIT Press, 1998. 57, 64

[126] J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics*, pages 647–656, Lancaster University, UK, Mar. 2003. 24, 41

[127] J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí. Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation*, 39(2-3) :123–164, 2005. 6

[128] J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. ISO-TimeML : An International Standard for Semantic Annotation. In N. C. C. Chair), K. Choukri, B. Maegaard,

J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). 6, 31, 33

[129] M. Richardson and P. Domingos. The intelligent surfer : Probabilistic combination of link and content information in pagerank. *Advances in Neural Information Processing Systems*, 2 :1441–1448, 2002. 88

[130] A. Rodrigo, A. Peñas, and F. Verdejo. Overview of the Answer Validation Exercise 2008. In ZCLEF08 [164]. 81

[131] A. Rodríguez, N. V. de Weghe, and P. D. Maeyer. Simplifying Sets of Events by Selecting Temporal Relations. In *Geographic Information Science, Third International Conference, GIScience 2004*, volume 3234/2004 of *Lecture Notes in Computer Science*, pages 269–284, Adelphi, MD, USA, Oct. 2004. Springer Berlin / Heidelberg. 40

[132] S. Rosset, O. Galibert, G. Adda, and E. Bilinski. The LIMSI Qast systems : comparison between human and automatic rules generation for question-answering on speech transcriptions. In *ASRU*, Kyoto, 2007. 79

[133] J. Rowley. The wisdom hierarchy : representations of the DIKW hierarchy. *Journal of Information Science*, 33(2) :163–180, 2007. 43

[134] I. Russo, T. Caselli, and F. Rubino. Recognizing deverbal events in context. In *Proceedings of 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011), poster session*, Tokyo, Japan, feb 2011. Springer. 24

[135] E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2012 Tweet Contextualization Track. In *Working Notes for the CLEF 2012 Workshop*, Rome (Italy), Sept. 2012. 100

[136] R. Saurí, R. Knippen, M. Verhagen, and J. Pustejovsky. Evita : A Robust Event Recognizer for QA Systems. In *Proceedings of HLT/EMNLP*, pages 700–707, Oct. 2005. 6

[137] G. Savova, S. Bethard, W. Styler, J. Martin, M. Palmer, J. Masanz, and W. Ward. Towards temporal relation discovery from the clinical narrative. In *AMIA Annu Symp Proc*, pages 568–572, 2009. 8

[138] S. Sekine. On-demand information extraction. In *Proceedings of the COLING/ACL (poster session)*, pages 731–738, Sydney, Australia, 2006. Association for Computational Linguistics. 12

[139] B. Settles. Closing the Loop : Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2011)*, pages 27–31, Edinburgh, UK, 2011. 12

[140] S. Sharoff. Creating General-Purpose Corpora Using Automated Search Engine Queries. In *WaCky ! Working papers on the Web as Corpus. Gedit*, 2006. 86

[141] Y. Shinyama and S. Sekine. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of HLT-NAACL*, pages 304–311, New York City, USA, 2006. 12

[142] J. Strötgen and M. Gertz. Temporal Tagging on Different Domains : Challenges, Strategies, and Gold Standards. In ZLREC2012 [167]. 6

[143] J. Strötgen and M. Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2) :269–298, 2013. 48, 63

[144] W. Sun, A. Rumshisky, and O. Uzuner. Evaluating temporal relations in clinical text : 2012 i2b2 Challenge. *Journal of the American Medical Information Association (JAMIA)*, apr 2013. 8, 35, 38

[145] L. Tanguy and N. Hathout. Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Actes de la 9ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2002)*, Nancy, France, 2002. 26

[146] X. Tannier and V. Moriceau. Building Event Threads out of Multiple News Articles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA, 2013. 45, 53, 57

[147] X. Tannier, V. Moriceau, B. Arnulphy, and R. He. Evolution of Event Designation in Media : Preliminary Study. In ZLREC2012 [167]. 27, 28

[148] X. Tannier and P. Muller. Evaluating Temporal Graphs Built from Texts via Transitive Reduction. *Journal of Artificial Intelligence Research*, 40 :375–413, 2011. 40

[149] D. Tribout. *Les conversions de nom à verbe et de verbe à nom.* PhD thesis, Université Paris Diderot - École doctorale de Sciences du Langage, 2010. 23

[150] P. D. Turney. Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. In ZACL02 [163], page 129–159. 71

[151] N. UzZaman and J. Allen. Event and Temporal Expression Extraction from Raw Text : First Step Towards a Temporally Aware System. *International Journal of Semantic Computing*, 2011. 6, 22

[152] N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, and J. Pustejovsky. SemEval-2013 Task 1 : TempEval-3 : Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings and the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, USA, jun 2013. ACL. 8, 22, 35

[153] A. Vallin, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the CLEF 2005 Multilingual Question Answering Track. In *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria, Sept. 2005. 82

[154] D. van de Velde. *Grammaire des événements.* Sens et structures. Presses Universitaires du Septentrion, 2006. 5

[155] Z. Vendler. Verbs and Times. *The Philosophical Review*, 66(2) :143–160, apr 1957. 23

[156] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. SemEval-2007 - 15 : TempEval Temporal Relation Identification. In *Proceedings of SemEval workshop at ACL 2007*, Prague, Czech Republic, June 2007. Association for Computational Linguistics, Morristown, NJ, USA. 8, 21, 22, 31

[157] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. SemEval-2010 - 13 : TempEval-2. In *Proceedings of SemEval workshop at ACL*, Uppsala, Sweden, 2010. 8, 21, 22

[158] W. Wang, R. Besançon, O. Ferret, and B. Grau. Filtering and clustering relations for unsupervised information extraction in open domain. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM 2011)*, CIKM '11, pages 1405–1414, Glasgow, Scotland, UK, 2011. ACM. 14

[159] W. Wang, R. et Cohen. Language-independent set expansion of named entities using the web. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 342–350, 2007. 87, 88

[160] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of of HLT-EMNLP-2005*, 2005. 63

[161] Y. Yang, T. Pierce, and J. G. Carbonell. A study on retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, Aug. 1998. ACM Press, New York City, NY, USA. 17, 46

[162] A. Yates. *Information Extraction from the Web : Techniques and Applications*. PhD thesis, University of Washington, 2007. 70

[163] *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics, Morristown, NJ, USA. 108, 111

[164] *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, Sept. 2008. 105, 109, 110

[165] G. Zhou, J. Su, J. Zhang, and M. Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, USA, 2005. Association for Computational Linguistics, Morristown, NJ, USA. 71

[166] L. Zhou, G. Melton, S. Parsons, and G. Hripcsak. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform*, 39(4) :424–439, 2006. 8

[167] *Proceedings of the Eighth International Language Resources and Evaluation (LREC'2012)*, Istanbul, Turkey, May 2012. 102, 103, 108, 110, 111

[168] P. Zweigenbaum and X. Tannier. Extraction des relations temporelles entre événements médicaux dans des comptes rendus hospitaliers. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2013, article court)*, Les Sables d'Olonne, France, June 2013. 35