Désambiguïsation d'entités pour l'induction non supervisée de schémas événementiels

Kiem-Hieu Nguyen[1, 2] Xavier Tannier[3, 1] Olivier Ferret[2] Romaric Besançon[2]
(1) LIMSI-CNRS
(2) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, F-91191, Gif-sur-Yvette
(3) Univ. Paris-Sud
{nguyen,xtannier}@limsi.fr, {olivier.ferret,romaric.besancon}@cea.fr

Résumé. Cet article présente un modèle génératif pour l'induction non supervisée d'événements. Les précédentes méthodes de la littérature utilisent uniquement les têtes des syntagmes pour représenter les entités. Pourtant, le groupe complet (par exemple, "un homme armé") apporte une information plus discriminante (que "homme"). Notre modèle tient compte de cette information et la représente dans la distribution des schémas d'événements. Nous montrons que ces relations jouent un rôle important dans l'estimation des paramètres, et qu'elles conduisent à des distributions plus cohérentes et plus discriminantes. Les résultats expérimentaux sur le corpus de MUC-4 confirment ces progrès.

Abstract.
Entity disambiguation for event template induction
In this paper, we present an approach for event induction with a generative model. This model makes possible to consider more relational information than previous models, and has been applied to noun attributes. By their influence on parameter estimation, this new information make probabilistic topic distribution more discriminative and more robust. We evaluated different versions of our model on MUC-4 datasets.

Mots-clés : Événements, modèle génératif, désambiguïsation d'entités, échantillonnage de Gibbs.

Keywords: Event Induction, Generative Model, Entity Disambiguation, Gibbs Sampling.

1 Introduction

L'extraction d'information s'est initialement définie et se définit toujours en grande partie au travers des tâches prescrites par les évaluations MUC (Message Understanding Conferences (Grishman & Sundheim, 1996)) et plus particulièrement sa tâche de remplissage de formulaire. Dans un tel contexte, l'objectif est d'extraire à partir de textes les événements d'un certain type ainsi que les éléments permettant de les caractériser. Le formulaire constitue la représentation du type d'événement considéré et par là même, la spécification des informations à extraire. Un formulaire rassemble donc l'ensemble de ces informations, prenant la forme dans un certain nombre de cas d'entités nommées, en précisant le rôle occupé par chaque information vis-à-vis de l'événement auquel elle se rattache. Pour un événement comme un tremblement de terre par exemple, le formulaire regroupera typiquement des informations (rôles) comme sa localisation, sa date, sa magnitude et les dégâts qu'il a pu occasionner (Jean-Louis *et al.*, 2011).

Malgré les efforts entrepris pour définir des tâches génériques, comme la reconnaissance d'entités nommées, la tâche de remplissage de formulaire reste très dépendante du type d'événement considéré. Ainsi, le travail réalisé pour développer un système concernant un type d'événement donné est pour une bonne part à recommencer pour un autre type d'événement. Quelques travaux se sont néanmoins attachés à limiter l'effort nécessaire à la définition d'un nouveau système d'extraction d'événements. Freedman *et al.* (2011) abordent la question par le biais d'une association entre des méthodes génériques fondées sur l'apprentissage et des règles définies manuellement. Grishman & He (2014) optent quant à eux pour la conjugaison d'un nombre restreint d'exemples fournis manuellement et de méthodes, en particulier distributionnelles, permettant d'étendre automatiquement la couverture de ces exemples.

D'autres travaux ont poussé plus loin cette logique en voulant offrir aux utilisateurs des modes d'extraction de l'infor-

mation plus souples et plus ouverts quant à la spécification de leur besoin informationnel. Ainsi, l'approche *On-demand information extraction* (Sekine, 2006), préfigurée dans Hasegawa *et al.* (2004) et concrétisée par la *Preemptive Information Extraction* (Shinyama & Sekine, 2006), vise à induire l'équivalent d'un formulaire à partir d'un ensemble de documents représentatifs des informations à extraire, documents typiquement obtenus par le biais de requêtes soumises à un moteur de recherche. Ce courant de recherche s'est ensuite davantage orienté vers l'extraction de relations, avec notamment Kathrin Eichler & Neumann (2008), Rosenfeld & Feldman (2007) et plus récemment Min *et al.* (2012), que vers l'extraction d'événements.

Dans cet article, nous nous plaçons dans la voie tracée initialement par Hasegawa *et al.* (2004) et Sekine (2006) en considérant la possibilité d'induire des représentations d'événements à partir de textes. Plus globalement, nous cherchons à construire une base de connaissances événementielles à partir de larges corpus journalistiques afin d'offrir des moyens d'accès structurés à ces corpus. Nous nous concentrons dans le cadre du travail présenté dans cet article sur le processus d'induction de schémas d'événements.

2 Objectif

Notre objectif global est de modéliser les événements décrits dans un corpus journalistique et d'identifier sans supervision les schémas (ou formulaires) récurrents ainsi que les rôles associés permettant de représenter ces événements. L'idée initiale de l'approche est de regrouper les entités [1] correspondant à certains rôles dans des événements en fonction de leurs relations avec les mêmes prédicats. Par exemple, dans un corpus sur des attentats terroristes, les mots qui sont objets des verbes *kill*, *attack* peuvent être regroupés et caractérisés par un rôle *VICTIM*. Le résultat de cette identification est donc constitué par un ensemble de groupes (*clusters*) contenant des mots et des relations associés à une probabilité d'appartenance à ce groupe (voir un exemple plus loin, Figure 4). Ces groupes ne sont pas nommés mais représentent chacun un rôle d'événement.

L'approche que nous proposons ici a pour objectif d'améliorer cette approche initiale en aidant à la désambiguïsation des entités. En effet, certaines entités ambiguës, comme "*man*" ou "*soldier*", peuvent correspondre à deux rôles différents (victime ou auteur de l'attaque). Une entité comme "*terrorist*" peut se retrouver mêlée aux victimes lorsque les articles relatent qu'un terroriste est tué par la police (et est donc également objet de *kill*). Notre hypothèse est que le contexte proche des entités est porteur d'information pour aider à leur désambiguïsation. Par exemple, le fait que l'entité "*man*" soit associée à "*armed*", "*dangerous*", "*heroic*" ou "*innocent*" peut conduire à une meilleure attribution et donc définition des rôles. Nous introduisons donc dans le modèle, en plus des relations avec les prédicats, les relations des entités avec leurs attributs (modifieurs syntaxiques).

Dans la pratique, nous utilisons un modèle génératif proche des "modèles de sujet" (*topic models*), mais sans modéliser la notion de document, habituellement centrale. Ces modèles permettent une catégorisation "douce" (*soft clustering*), c'est-à-dire que chaque mot et chaque relation peuvent apparaître dans plusieurs groupes. Ceci permet de traiter efficacement les nombreux mots et relations ambigus, qui peuvent selon le contexte tenir des rôles différents. En pratique, un modèle génératif considère que les observations (ici, les entités et les relations du corpus) peuvent être générées à partir des rôles. Pour cela, les rôles sont définis par des distributions probabilistes sur les prédicats, les entités et leurs arguments syntaxiques. L'enjeu est d'apprendre les paramètres de ces distributions (ici, par échantillonnage de Gibbs) pour que le résultat soit le plus proche possible des observations initiales (*maximum a posteriori* – MAP).

Pour évaluer la qualité de notre approche, nous comparons les groupes de mots produits avec des formulaires et des rôles de référence, qui sont pour leur part nommés et contiennent des mots du corpus associés aux phrases dans lesquelles ils apparaissent. Pour effectuer cette comparaison, nous utilisons une stratégie automatique et empirique d'association entre les rôles du système et ceux de la référence, de façon similaire aux travaux précédents dans le domaine.

N.B. : la terminologie anglophone pour le domaine considéré étant plus répandue et plus stabilisée, nous précisons les termes anglais correspondant aux termes français que nous avons choisis : formulaire : *template* ; rôle : *slot* ; modèle de sujet : *topic model* ; relation-prédicat : parfois nommée *event trigger*, ou *verb path*, ou tout simplement *relation*.

Après une brève présentation des travaux reliés (Section 3), nous précisons la représentation que nous avons choisie pour les entités et les relations que nous extrayons (Section 4). Nous décrivons ensuite le modèle génératif mis en œuvre pour le regroupement en rôles (Section 5), avant de proposer plusieurs expérimentations et évaluations (Section 6).

1. Dans la pratique, tout groupe nominal est une entité candidate pour un rôle.

3 Travaux reliés

La problématique de l'induction de schémas d'événements à partir de textes n'est pas nouvelle. Elle plonge en effet ses racines dans les travaux sur l'acquisition des schémas (Lebowitz, 1983) utilisés par les systèmes de compréhension de textes de la fin des années 70 et du début des années 80 (DeJong, 1982). Cette même perspective se retrouve dans Ferret & Grau (1997). Une des premières introductions de cette problématique pour la création automatique de templates dans le domaine de l'extraction d'information est le fait de Collier (1998). Elle apparaît ensuite dans Harabagiu (2004) sous la forme de la structuration de thèmes événementiels tels que ceux considérés dans les évaluations *Topic Detection and Tracking* (Wayne, 1998). Les schémas ainsi formés ont ensuite été utilisés à la fois en extraction d'information et en résumé automatique. L'intérêt pour cette problématique s'est aussi étendu au domaine du question-réponse par la volonté de créer automatiquement une représentation des événements à partir de textes (Filatova *et al.*, 2006) pour améliorer la recherche de réponse à des questions événementielles (Filatova, 2008). L'essor plus récent des approches faiblement supervisées a enfin vu le développement de plusieurs travaux importants abordant le sujet selon différents angles. Qiu *et al.* (2008) l'ont ainsi envisagé comme un problème de clustering dans un graphe de propositions construit à partir des textes. Regneri *et al.* (2010) l'ont abordé comme un problème d'alignement de séquences d'événements. Bejan (2008) a adopté pour sa part une approche générative fondée sur le paradigme de l'allocation de Dirichlet latente (LDA) en considérant qu'un document est représenté comme une distribution de probabilité sur un ensemble de schémas d'événements et que chacun de ces schémas est lui-même défini comme une distribution de probabilité sur un ensemble de frames sémantiques de type FrameNet (Baker *et al.*, 1998). Des modèles génératifs plus spécifiques au problème considéré que la simple transposition d'une approche LDA ont été ensuite proposés dans des travaux comme Chambers (2013), Cheung *et al.* (2013) et dernièrement Frermann *et al.* (2014). Enfin, Chambers & Jurafsky (2008), Chambers & Jurafsky (2009), Chambers & Jurafsky (2011), amélioré par Balasubramanian *et al.* (2013), et Chambers (2013) se sont plus particulièrement focalisés sur l'émergence de rôles et la découverte de chaînes d'événements pour induire des schémas narratifs à partir de textes en se fondant sur la résolution de coréférence et en prenant en compte la dimension temporelle pour l'ordonnancement des événements au sein de ces schémas. Le travail que nous présentons dans cet article se situe dans le prolongement de Chambers (2013). Nous aurons d'ailleurs l'occasion de décrire son approche plus en détails et de la comparer à la nôtre en termes qualitatifs et quantitatifs dans la Section 6.

4 Représentation des entités et des relations

Une entité est représentée par un triplet comprenant un mot (la tête de l'entité), une liste de "relations-attributs" et une liste de "relations-prédicats". Considérons l'exemple suivant :

(1) Two armed men attacked the police station and killed a policeman. An innocent young man was also wounded.

Comme illustré par la Figure 1, quatre entités sont distinguées, représentées par quatre triplets. Les têtes d'entités sont extraites des syntagmes.

Une relation-prédicat est composée d'un prédicat (*attack*, *kill*, *explosion*) et d'un type de dépendance (sujet – *nsubj*, object – *dobj*, etc.). Une relation-attribut est composée d'un argument (*armed*, *police*, *young*) et d'un type de dépendance (modifieur adjectival – *amod*, nominal – *nn* ou verbal – *vmod*). Notons que le mot de l'entité gouverne la relation-attribut mais est gouverné dans la relation-prédicat. Nous utilisons l'analyseur de Stanford (Manning *et al.*, 2014) pour l'analyse syntaxique et la résolution des coréférences.

Une *entité* est extraite pour chaque groupe nominal dont la tête (nom commun ou propre) est liée à au moins un prédicat. Les pronoms ne sont pas considérés. Une *relation-prédicat* est quant à elle extraite pour chaque verbe ou nom d'événe-

Triplets	Relations-attributs	Tête	Relations-prédicats
#1	armed:amod	man	[attack:nsubj, kill:nsubj]
#2	police:nn	station	attack:dobj
#3	-	policeman	kill:dobj
#4	[innocent:amod, young:amod]	man	wound:dobj

FIGURE 1 – Représentation des entités comme des triplets de ([relations-attributs], tête, [relations-prédicats])

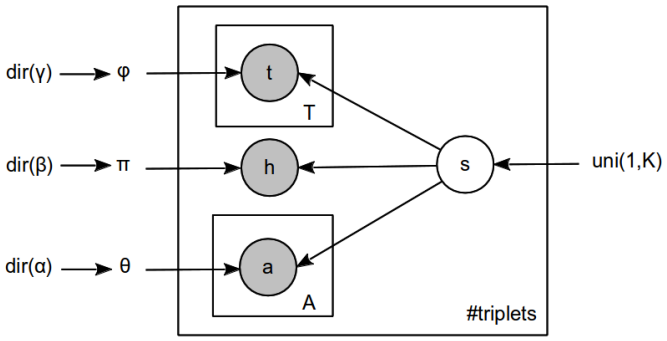


FIGURE 2 – Modèle génératif pour l'induction des formulaires d'événements

ment lié à une entité. Les noms d'événements sont les noms entrant dans les catégories *noun.EVENT* et *noun.ACT* dans WordNet (Miller, 1995). Une entité peut être liée à plus d'une relation-prédicat. Cette multiplicité peut avoir une origine intra-phrastique, comme c'est le cas avec la coordination présente dans la première phrase de notre exemple, ou inter-phrastique dans le cas des coréférences. Nous fusionnons en effet les différentes mentions d'une chaîne de coréférence en une seule entité, à laquelle sont rattachées les relations-prédicats de chaque mention. Il serait même possible de fusionner des mentions inter-document si un système traitait ce type de coréférences avec une précision satisfaisante. Enfin, une *relation-attribut* est extraite pour chaque adjectif, nom ou verbe jouant le rôle de modifieur adjectival, nominal ou verbal d'un nom. Si plusieurs modifieurs existent, seul l'élément le plus proche de la première mention de l'entité est conservé. Cette heuristique permet de ne pas conduire à une influence supérieure des attributs par rapport aux relations-prédicats (les attributs multiples étant plus fréquents). Elle est appropriée pour la langue anglaise, où elle permet d'omettre un grand nombre d'adjectifs non discriminants pour l'entité, mais pourrait être rediscutée pour une autre langue. Les expériences finales montrent de meilleures performances lorsque l'on se limite à un attribut.

5 Modèle génératif

Nous présentons dans cette section le modèle de sujet que nous avons mis en œuvre, avec une description algorithmique et graphique de son processus de génération puis un détail de l'estimation de ses paramètres.

5.1 Description du modèle

La représentation graphique de notre modèle est présentée à la Figure 2. Pour chaque triplet représentant une entité e, le modèle choisit un rôle s pour cette entité à partir d'une distribution uniforme $uni(1, K)$, où K est le nombre de rôles. La tête h du syntagme de l'entité est ensuite générée à partir d'une distribution multinomiale π_s. Chaque relation-prédicat t de l'ensemble T_e des relations-prédicats est générée à partir d'une distribution multinomiale ϕ_s. Enfin, chaque relation-attribut a de l'ensemble A_e des relations-attributs est également générée à partir d'une distribution multinomiale θ_s. Les distributions θ, π et ϕ sont générées par les lois de Dirichlet $dir(\alpha)$, $dir(\beta)$ et $dir(\gamma)$, respectivement.

Étant donné un ensemble d'entités E, notre modèle (π, ϕ, θ) est défini par :

$$P_{\pi,\phi,\theta}(E) = \prod_{e \in E} P_{\pi,\phi,\theta}(e) \tag{2}$$

sans faire de distinction entre les documents contenant les entités. La probabilité de chaque entité e est définie par :

$$\begin{aligned} P_{\pi,\phi,\theta}(e) &= P(s) \\ &\times P(h|s) \\ &\times \prod_{t \in T_e} P(t|s) \\ &\times \prod_{a \in A_e} P(a|s) \end{aligned} \quad (3)$$

Le processus génératif est le suivant :

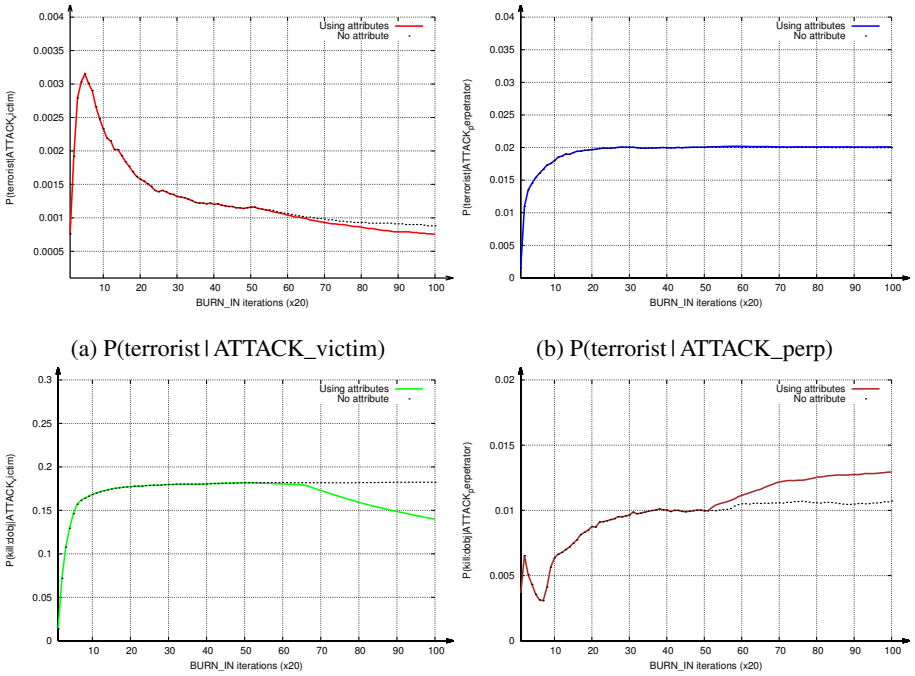
for *rôle* $s \leftarrow 1$ **to** K **do**
  Générer une distribution de têtes d'entités π_s à partir d'une loi de Dirichlet $dir(\beta)$;
  Générer une distribution de relations-attributs θ_s à partir d'une loi de Dirichlet $dir(\alpha)$;
  Générer une distribution de relations-prédicats ϕ_s à partir d'une loi de Dirichlet $dir(\gamma)$;
end
for *entité* $e \in E$ **do**
  Générer un rôle s à partir d'une distribution uniforme $uni(1, K)$;
  Générer une tête h à partir d'une distribution multinomiale π_s ;
  for $i \leftarrow 1$ **to** $|T_e|$ **do**
    Générer une relation-prédicat t_i à partir d'une distribution multinomiale ϕ_s;
  end
  for $j \leftarrow 1$ **to** $|A_e|$ **do**
    Générer une relation-attribut a_j à partir d'une distribution multinomiale θ_s;
  end
end

5.2 Apprentissage des paramètres

L'estimation des paramètres du modèle est effectuée par la méthode d'échantillonnage de Gibbs (Griffiths, 2002). La variable s du rôle est échantillonnée par intégration de toutes les autres variables du modèle. Les modèles précédents (Cheung *et al.*, 2013; Chambers, 2013) sont fondés sur une modélisation des sujets au niveau du document, dans la tradition des modèles comme l'allocation de Dirichlet latente (*Latent Dirichlet Allocation*, LDA (Blei *et al.*, 2003)). Notre modèle s'affranchit de cette notion de document puisque ce niveau de structure n'a finalement pas d'impact sur les rôles qu'il contient. L'entrée du modèle est donc une chaîne continue de triplets d'entités. Les frontières du document sont uniquement utilisées à l'issue de l'apprentissage pour l'étape de filtrage (décrite à la Section 6.5). La distribution des rôles est donc globale et non spécifique à chaque document, ce qui est plus en phase avec l'induction de structures à l'échelle du corpus. De plus, la distribution *a priori* des rôles est ignorée en initialisant avec une distribution uniforme, cas particulier d'une distribution de Bernouilli généralisée.

L'attribution des rôles par échantillonnage dépend des états initiaux et du hasard. Dans notre implémentation de l'échantillonnage de Gibbs, nous utilisons 2 000 itérations de *burn in* sur un total de 10 000 itérations. Cette étape de *burn in* permet de s'assurer que les paramètres convergent vers un état stable avant d'utiliser les échantillons pour l'estimation des distributions de probabilités. De plus, après le *burn in*, un intervalle de 100 itérations est appliqué entre deux estimations des paramètres pour éviter une trop grande proximité entre deux échantillons successifs.

Enfin, les attributs étant moins porteurs de sens mais venant plutôt "en support", la phase de *burn in* ne les considère pas et n'échantillonne que les entités et les prédicats. L'état stable obtenu est ensuite utilisé comme initialisation pour un échantillonnage sur les trois éléments. Ainsi, les éléments non ambigus sont peu affectés par les attributs tandis que les éléments plus "sensibles", comme par exemple les entités ambiguës "*man*", "*soldier*" ou les relations "*kill:dobj*" et "*attack:nsubj*" voient leurs probabilités modifiées. Sans les attributs, les "terroristes" ("*guerrilla*", "*terrorists*", etc.) sont souvent mélangés aux victimes, car souvent "tués" (c'est-à-dire, en relation avec *kill:dobj*) par la police. Grâce à l'ajout des adjectifs et des attributs, la séparation entre les "*perpetrators*" (qui sont par exemple "*armed*" et "*dangerous*") et les "*victims*" (plutôt "*heroic*" ou "*innocent*") est facilitée. On voit par exemple à la Figure 3, qui montre l'évolution des probabilités de certains éléments à mesure des itérations de *burn in*, que la probabilité de la relation "*kill:dobj*" diminue dans le rôle correspondant aux victimes, de même que celle de l'entité "*terrorist*" dans ce même rôle.



(a) P(terrorist | ATTACK_victim) (b) P(terrorist | ATTACK_perp)

(c) P(kill:dobj | ATTACK_victim) (d) P(kill:dobj | ATTACK_perp)

FIGURE 3 – Les 2 000 premières itérations (*burn in*) de l'échantillonnage pour 4 rôles, avec les attributs ajoutés à mi-parcours (50 %, ligne continue), ou sans les attributs (ligne pointillée)

6 Évaluation

Pour l'évaluation, nous utilisons le corpus MUC-4 (Message Understanding Conference (Sundheim, 1991)) afin de nous comparer aux autres systèmes dont les résultats sont généralement obtenus avec ce corpus. Nous utilisons les métriques traditionnelles de précision, rappel et F-mesure. Dans ce qui suit, nous présentons d'abord le corpus utilisé (Section 6.1), puis les paramètres du système (6.2), avant de détailler la méthode d'association entre les rôles appris par le système et les rôles du corpus de référence (6.3). Par la suite, nous présentons une première expérience (Section 6.4) montrant l'intérêt d'utiliser les attributs des entités pour la modélisation des rôles. La seconde expérience (Section 6.5) étudie l'impact d'une étape de classification des documents. Enfin, nous comparons nos résultats avec ceux faisant référence, et plus particulièrement à Chambers (2013), tant sur des aspects quantitatifs que qualitatifs (Section 6.6).

6.1 Corpus

Le corpus MUC-4 contient 1 700 articles journalistiques en langue anglaise concernant des incidents et des attaques terroristes en Amérique Latine. Le corpus est divisé en 1 300 documents pour le développement et quatre ensembles de test contenant chacun 100 documents. Nous suivons les règles établies par les travaux antérieurs pour garantir la comparabilité de nos résultats avec ces travaux (Patwardhan & Riloff, 2007; Chambers & Jurafsky, 2011). Quatre types de formulaires sont présents : *Arson*, *Attack*, *Bombing*, et *Kidnapping* ; pour chacun de ces formulaires, quatre rôles peuvent être pertinents : *Instrument*, *Target*, *Victim*, *Perpetrator* (fusion de *Perpetrator_Individual* et *Perpetrator_Organization*). L'association entre les réponses du système et la référence est fondée sur l'association entre les têtes des groupes nominaux uniquement, la tête étant définie comme le nom le plus à droite du syntagme, ou comme le dernier nom avant une préposition "of". Certains formulaires et rôles dit "optionnels" sont ignorés lors du calcul du rappel. Les types des formulaires sont ignorés pour l'évaluation, ce qui signifie que le *perpetrator* d'un *bombing* et celui d'un *attack* au niveau de la référence peuvent se retrouver associés dans un même rôle induit.

BOMBING_instrument		
Attributs	**Têtes**	**Relations avec prédicats**
car:nn	bomb	explode:nsubj
powerful:amod	fire	hear:dobj
explosive:amod	explosion	place:dobj
dynamite:nn	blow	cause:nsubj
heavy:amod	charge	set:dobj
KIDNAPPING_victim		
Attributs	**Têtes**	**Relations avec prédicats**
several:amod	people	arrest:dobj
other:amod	person	kidnap:dobj
responsible:amod	man	release:dobj
military:amod	member	kill:dobj
young:amod	leader	identify:prep_as

FIGURE 4 – Distributions apprises par le modèle que nous nommerons plus tard *HT+A* pour les rôles *BOMBING_instrument* et *KIDNAPPING_victim* (les noms des rôles étant attribués après l'association décrite à la Section 6.3)

6.2 Paramètres du système

Les hyperparamètres [2] sont fixés selon les meilleurs résultats obtenus sur l'ensemble de développement. Le nombre de rôles est ainsi fixé à $s = 35$. Les paramètres *a priori* des lois de Dirichlet sont fixés à $\alpha = 0,1$, $\beta = 1$ et $\gamma = 0,1$. Le modèle est appris sur l'ensemble des données. L'association des rôles (Section 6.3) est effectuée sur les deux premiers ensembles de test. Les sorties des deux derniers ensembles de test sont ensuite évaluées suivant cette association. Par ailleurs, l'échantillonnage étant fondé sur des tirages aléatoires, les valeurs de précision, rappel et F-mesure indiquées sont des moyennes sur 10 exécutions du système.

6.3 Association des rôles du système avec le corpus de référence

Notre modèle apprend K rôles et assigne chaque entité d'un document à l'un de ces rôles. L'association des rôles consiste à faire correspondre chaque rôle de la référence à un rôle équivalent appris par le modèle. Notons que parmi les K rôles appris, certains sont peu pertinents, et d'autres, parfois de très bonne qualité, contiennent des entités ne faisant pas partie de la référence (informations spatio-temporelles, contexte des protagonistes, etc.). Il n'est pas donc illogique que le nombre de rôles appris soit très supérieur au nombre de rôles attendus pour un événement.

De la même façon que les travaux précédents sur le même type de tâche, nous effectuons l'association des rôles de façon automatique et empirique. Nous réservons une partie du corpus de développement à l'opération. Chaque rôle de référence est associé au rôle appris par le système qui permet d'obtenir la meilleure F-mesure. Ici, deux rôles ayant le même nom mais appartenant à deux formulaires différents sont considérés de façon distincte.

À titre d'exemple, la Figure 4 montre les mots obtenant le poids le plus élevé pour les rôles *BOMBING_instrument* et *KIDNAPPING_victim*. L'association ainsi créée est conservée pour l'application sur le corpus de test.

6.4 Utilisation des attributs des entités

Deux versions différentes de notre modèle sont évaluées : *HT* utilise uniquement les têtes des entités et leurs relations avec des événements tandis que *HT+A* introduit également la distribution des attributs des entités.

Pour estimer plus extensivement le gain du modèle avec les attributs, nous avons également fait varier la richesse de l'entrée du modèle, notamment pour tester les éventuelles interactions entre les attributs et l'utilisation des informations

2. Tandis que les *paramètres* du système sont des variables à apprendre par le système, les *hyperparamètres* sont des constantes fixées manuellement par l'expérimentateur.

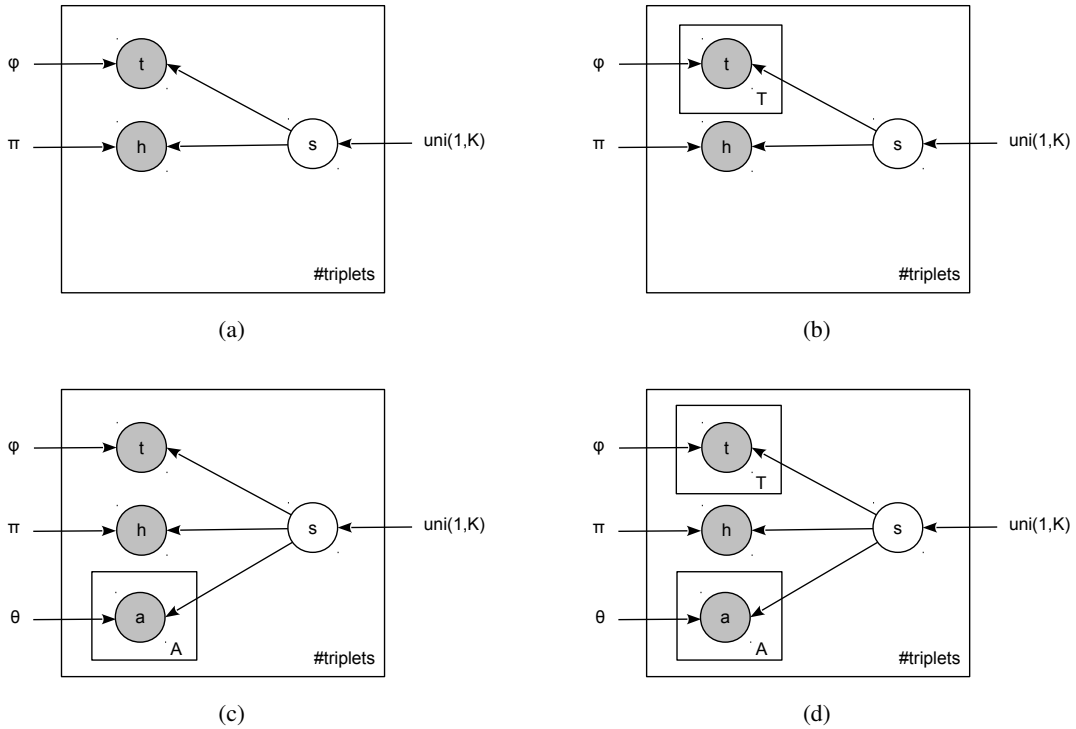


FIGURE 5 – Variations du modèle (les distributions de Dirichlet sont omises) : 5a) modèle HT avec entrée simple. Ce modèle est équivalent à 5b) si on n'a qu'un seul prédicat lié à chaque entité (T=1) ; 5b) modèle HT avec entrée multiple ; 5c) modèle HT+A sur entrée simple ; 5d) modèle HT+A sur entrée multiple

Data	HT			HT+A		
	P	R	F	P	R	F
Single	29,59	51,17	37,48	30,22	52,41	38,33
Multiple	29,32	52,21	37,52	30,82	51,68	38,55
Coref	39,99	53,53	40,01	32,42	54,59	40,62

TABLE 1 – L'amélioration apportée par l'utilisation des attributs d'entités

de coréférence. L'entrée dite "simple" ne permet de relier qu'un seul événement à chaque entité. Dans cette configuration, l'énoncé "*an armed man attacked the police station and killed a policeman*" produit deux triplets pour l'entité "man" : *(armed:amod, man, attack:nsubj)* et *(armed:amod, man, kill:nsubj)*. Dans l'entrée "multiple", cette duplication de l'entité n'est pas nécessaire puisque plusieurs relations peuvent être reliées à une même entité, conduisant par exemple à *(armed:amod, man, [attack:nsubj, kill:nsubj])*. Enfin, la sortie appelée "coref" ajoute au modèle précédent la prise en compte des relations multiples provenant d'une relation de coréférence détectée par l'analyseur. Par exemple, si "an armed man" est l'antécédent de "he" dans "*He was arrested three hours later*", alors l'entrée du modèle sera *(armed:amod, man, [attack:nsubj, kill:nsubj, arrest:dobj])*.

Notons que ces différences dans les données fournies au modèle ont une influence sur l'organisation du modèle lui-même. Le Figure 5 montre les variations de ce modèle selon que l'entrée est "simple" ou "multiple".

La Table 1 montre une amélioration systématique apportée par l'utilisation des attributs, que ce soit avec ou sans la résolution de coréférence. La meilleure performance (F-mesure de 40,62) est obtenue par le modèle le plus complexe. Ainsi, l'ajout commun des attributs et de la coréférence conduit à un gain de 3 points de F-mesure.

6.5 Classification des documents

Dans cette seconde expérience, nous avons ajouté à notre modèle une étape de classification des documents après la phase d'apprentissage.

Le corpus MUC-4 contient un grand nombre de documents "non pertinents", c'est-à-dire ne comprenant aucune annotation selon les formulaires décrits plus haut. 567 documents sur les 1 300 de l'ensemble de développement sont non pertinents. Ils sont pourtant difficiles à éliminer automatiquement car ils contiennent de nombreuses allusions au terrorisme, avec les mots "bomb", "force", "guerrilla" et bien d'autres. Dans notre modèle comme dans les précédents, l'annotation erronée d'entités dans ces documents est la cause d'un grand nombre de faux positifs. Le but de notre classification de documents *a posteriori* est donc d'augmenter la précision en limitant la réduction du rappel.

Étant donné un document d dont les entités ont été assignées à des rôles du modèle, nous devons décider si ce document est pertinent ou pas. Nous définissons le score de pertinence du document de la façon suivante :

$$pertinence(d) = \frac{\sum_{e \in d : s_e \in S_m} \sum_{t \in T_e} P(t|s_e)}{\sum_{e \in d} \sum_{t \in T_e} P(t|s_e)} \quad (4)$$

où e est une entité dans le document d ; s_e (de la liste S_m de tous les rôles) est le rôle attribué à e ; t est un prédicat de la liste T_e de tous les prédicats (rappelons que chaque entité et chaque prédicat possède une probabilité pour chaque rôle).

L'équation (4) définit donc le score d'une entité comme la somme des probabilités conditionnelles des prédicats qui lui sont associés, dans le rôle qui lui a été assigné. Le score de pertinence du document est alors proportionnel au score des entités associés aux rôles qui ont été sélectionnés. Si le score d'un document est supérieur à un certain seuil λ, alors le document est considéré comme pertinent. La valeur de λ est fixée à $\lambda = 0,02$, valeur qui maximise la F-mesure sur l'ensemble de développement.

La Table 2 nous montre l'amélioration apportée par l'application de la classification des documents. La précision augmente nettement et le rappel diminue peu, conduisant à une hausse de la F-mesure. Nous nous comparons également avec un classifieur "oracle", virtuel, qui filtrerait parfaitement les documents non pertinents en conservant les pertinents. Les performances de notre modèle avec ce classifieur oracle montrent qu'une marge d'amélioration non négligeable existe encore pour exploiter la classification des documents.

Le filtrage des éléments non pertinents se retrouve aussi bien au niveau des méthodes d'extraction d'information non supervisées que supervisées. Dans ce dernier cas, il est souvent appliqué à un niveau plus fin que le document, typiquement la phrase (Patwardhan & Riloff, 2009, 2007). Pour ce qui est des méthodes non supervisées, Chambers (2013) réalise ce filtrage de façon indirecte par la prise en compte des documents au sein de son modèle de sujet tandis que Chambers & Jurafsky (2011) et Cheung *et al.* (2013) utilisent comme nous les *clusters* appris pour classifier les documents. La pertinence d'un document est estimée dans ce cas à partir de statistiques sur les prédicats.

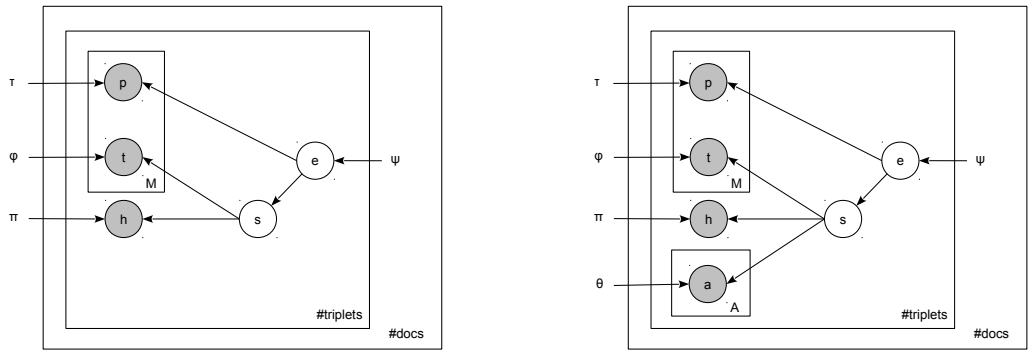
6.6 Comparaison avec l'état de l'art

Nous avons enfin comparé nos résultats avec ceux des systèmes d'extraction non supervisée de formulaires de référence. Nous avons en particulier ré-implémenté la méthode proposée par Chambers (2013), en ajoutant également à son modèle une distribution des attributs comparable à la nôtre (cf. Figure 6).

Les différences principales entre ce modèle et le nôtre sont les suivantes :

	P	R	F
HT + A	32,42	54,59	40,62
HT + A + classification	35,57	53,89	42,79
HT + A + classification oracle	44,58	54,59	49,08

TABLE 2 – Amélioration apportée par la classification des documents



(a) (b)

FIGURE 6 – Modèle de Chambers (2013) : 6a) version originale ; 6b) avec une distribution supplémentaire pour les attributs

Système	P	R	F
(Cheung *et al.*, 2013)	32	37	34
(Chambers & Jurafsky, 2011)	**48**	25	33
Nest (Chambers, 2013)	41	41	41
Nest (réimplementé)	39	43	41
Nest (réimplementé) + Attributs	39	44	41
HT + A + doc. classification	36	**54**	**43**

TABLE 3 – Comparaison avec les systèmes non-supervisés état-de-l'art

1. Chambers (2013) ajoute une distribution ψ reliant les événements aux documents. Celle-ci rend le modèle plus complexe et peut-être moins intuitif puisque la notion de document ne devrait pas être *a priori* en lien avec les formulaires et les rôles : un document contient possiblement des références à plusieurs formulaires et l'association entre entités et rôles ne dépend pas du niveau du document. Cependant, cette distribution permet d'éviter le recours à une classification des documents séparée.
2. Chaque entité est liée à une variable d'événement p. Cet événement génère un prédicat pour chaque mention d'entité (rappelons que les mentions d'une entité sont toutes les occurrences d'une entité dans un document, y compris dans une chaîne de coréférence). Notre travail se concentre davantage sur le fait de prendre en compte une certaine pluralité des informations liées aux entités. Les relations-prédicats et les relations-attributs multiples sont ainsi traitées de la même façon. Cette multiplicité n'a pas seulement pour origine les chaînes de coréférence mais résulte aussi de la multiplicité "naturelle" des relations attachées à une entité (qui pourraient être des relations autres que syntaxiques). En conséquence, il est possible d'avancer que notre modèle offre une plus souplesse d'extension que celui de Chambers (2013), à la fois en termes de modélisation et de données d'entrée.
3. Enfin, Chambers (2013) ajoute une contrainte heuristique lors de l'échantillonnage imposant que le sujet et l'objet d'un même prédicat (par exemple, *kill:nsubj* et *kill:dobj*) ne soient pas distribués dans le même rôle. Notre modèle ne nécessite pas une telle heuristique.

Certains aspects concernant le prétraitement des données et les paramètres du modèle n'étant pas totalement précisés (Chambers, 2013), notre implémentation (menée sur les mêmes données) conduit à des résultats légèrement différents de ceux publiés. C'est pourquoi nous présentons ici les deux informations.

La Table 3 montre que notre modèle dépasse tous les autres modèles considérés, en particulier par un rappel bien supérieur, conduisant à une meilleure F-mesure. Il est également intéressant de constater que la prise en compte des attributs améliore le modèle de Chambers (2013), illustrant en cela la généralité de l'idée sous-jacente.

7 Conclusion et perspectives

Nous avons présenté dans cet article un modèle génératif permettant de modéliser les rôles tenus par des entités dans des schémas d'événements. Nous avons mis l'accent sur l'utilisation du contexte des entités protagonistes et nous avons proposé un modèle à la fois plus simple et plus efficace que ceux de l'état de l'art. Nous avons évalué ce modèle en comparant les rôles obtenus à ceux définis pour le corpus MUC-4. Une évaluation sur d'autres langues disposant d'analyseurs syntaxiques performants, comme le français, serait intéressante. Cependant, les corpus associés restent à construire.

Même si les résultats sont meilleurs que ceux obtenus précédemment, l'approche non supervisée est encore loin des résultats présentés par les approches supervisées. Les pistes d'amélioration sont donc nombreuses. En premier lieu, les caractéristiques du corpus MUC-4 sont un facteur limitant. Il est de petite taille et les rôles sont presque les mêmes pour chaque formulaire d'événement, ce qui n'est pas représentatif de la réalité. Un corpus de plus grande taille, même partiellement annoté, et présentant des schémas d'événements plus variés, permettrait d'envisager d'autres approches.

Comme nous l'avons montré, notre modèle a l'avantage de permettre une représentation unifiée de tous les types de relations ; une extension possible de notre travail serait donc de multiplier les relations (syntaxiques, sémantiques, thématiques, etc.) pour affiner les groupes générés.

Enfin, le protocole d'évaluation, devenu une sorte de standard de fait, est très imparfait. Notamment, la façon de réaliser l'alignement final avec les rôles de la référence peut influer fortement sur les résultats.

Remerciements

Ce projet a été partiellement financé par la Fondation de Coopération Scientifique "Campus Paris-Saclay" à travers le projet Digiteo ASTRE No 2013-0774D.

Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of ACL-COLING 1998*, p. 86–90, Montréal, Quebec, Canada.

BALASUBRAMANIAN N., SODERLAND S., MAUSAM & ETZIONI O. (2013). Generating Coherent Event Schemas at Scale. In *2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA.

BEJAN C. A. (2008). Unsupervised Discovery of Event Scenarios from Texts. In *Twenty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS 2008)*, p. 124–129, Coconut Grove, Florida.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

CHAMBERS N. (2013). Event Schema Induction with a Probabilistic Entity-Driven Model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, p. 1797–1807, Seattle, USA.

CHAMBERS N. & JURAFSKY D. (2008). Unsupervised Learning of Narrative Event Chains. In *ACL-08 : HLT*, p. 789–797, Columbus, Ohio.

CHAMBERS N. & JURAFSKY D. (2009). Unsupervised Learning of Narrative Schemas and their Participants. In *Proceedings of ACL-IJCNLP 2009*, p. 602–610, Suntec, Singapore.

CHAMBERS N. & JURAFSKY D. (2011). Template-Based Information Extraction without the Templates. In *49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL 2011)*, p. 976–986, Portland, Oregon, USA.

CHEUNG K. J. C., POON H. & VANDERWENDE L. (2013). Probabilistic Frame Induction. In *Proceedings of NAACL-HLT 2013*, p. 837–846.

COLLIER R. (1998). *Automatic Template Creation for Information Extraction*. PhD thesis, University of Sheffield.

DEJONG G. (1982). An overview of the FRUMP system. In W. LEHNERT & M. RINGLE, Eds., *Strategies for natural language processing*, p. 149–176. Lawrence Erlbaum Associates.

FERRET O. & GRAU B. (1997). An Aggregation Procedure for Building Episodic Memory. In *15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, p. 280–285, Nagoya, Japan.

FILATOVA E. (2008). *Unsupervised Relation Learning for Event-Focused Question-Answering and Domain Modelling*. PhD thesis, Columbia University.

FILATOVA E., HATZIVASSILOGLOU V. & MCKEOWN K. (2006). Automatic Creation of Domain Templates. In *COLING-ACL 2006*, p. 207–214, Sydney, Australia.

FREEDMAN M., RAMSHAW L., BOSCHEE E., GABBARD R., KRATKIEWICZ G., WARD N. & WEISCHEDEL R. (2011). Extreme Extraction – Machine Reading in a Week. In *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, p. 1437–1446, Edinburgh, Scotland, UK.

FRERMANN L., TITOV I. & PINKAL M. (2014). A Hierarchical Bayesian Model for Unsupervised Induction of Script Knowledge. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, p. 49–57, Gothenburg, Sweden.

GRIFFITHS T. (2002). *Gibbs sampling in the generative model of Latent Dirichlet Allocation*. Rapport interne, Stanford University.

GRISHMAN R. & HE Y. (2014). An Information Extraction Customizer. In P. SOJKA, A. HORÁK, I. KOPEČEK & K. PALA, Eds., *17th International Conference on Text, Speech and Dialogue (TSD 2014)*, volume 8655 of *Lecture Notes in Computer Science*, p. 3–10. Springer International Publishing.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference-6 : A Brief History. In *16^{th} International Conference on Computational linguistics (COLING'96)*, p. 466–471, Copenhagen, Denmark.

HARABAGIU S. (2004). Incremental Topic Representation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, Geneva, Switzerland.

HASEGAWA T., SEKINE S. & GRISHMAN R. (2004). Discovering Relations among Named Entities from Large Corpora. In *42^{nd} Meeting of the Association for Computational Linguistics (ACL'04)*, p. 415–422, Barcelona, Spain.

JEAN-LOUIS L., BESANÇON R. & FERRET O. (2011). Text Segmentation and Graph-based Method for Template Filling in Information Extraction. In *5^{th} International Joint Conference on Natural Language Processing (IJCNLP 2011)*, p. 723–731, Chiang Mai, Thailand.

KATHRIN EICHLER H. H. & NEUMANN G. (2008). Unsupervised Relation Extraction From Web Documents. In *6^{th} Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

LEBOWITZ M. (1983). Generalization from natural language text. *Cognitive Science*, **7**(1), 1–40.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 55–60, Baltimore, USA.

MILLER G. A. (1995). WordNet : a lexical database for English. *Communication of the ACM*, **38**(11), 39–41.

MIN B., SHI S., GRISHMAN R. & LIN C.-Y. (2012). Ensemble Semantics for Large-scale Unsupervised Relation Extraction. In *2012 Joint Conference EMNLP-CoNLL*, p. 1027–1037, Jeju Island, Korea.

PATWARDHAN S. & RILOFF E. (2007). Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In *Proceedings of the 2007 Joint Conference EMNLP-CoNLL*, p. 717–727, Prague, Czech Republic.

PATWARDHAN S. & RILOFF E. (2009). A Unified Model of Phrasal and Sentential Evidence for Information Extraction. In *Proceedings of EMNLP 2009*, p. 151–160, Singapore.

QIU L., KAN M.-Y. & CHUA T.-S. (2008). Modeling Context in Scenario Template Creation. In *Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, p. 157–164, Hyderabad, India.

REGNERI M., KOLLER A. & PINKAL M. (2010). Learning Script Knowledge with Web Experiments. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, p. 979–988, Uppsala, Sweden.

ROSENFELD B. & FELDMAN R. (2007). Clustering for unsupervised relation identification. In *Sixteenth ACM conference on Conference on information and knowledge management (CIKM'07)*, p. 411–418, Lisbon, Portugal.

SEKINE S. (2006). On-demand information extraction. In *Proceedings of COLING-ACL 2006*, p. 731–738, Sydney, Australia.

SHINYAMA Y. & SEKINE S. (2006). Preemptive Information Extraction using Unrestricted Relation Discovery. In *HLT-NAACL 2006*, p. 304–311, New York City, USA.

SUNDHEIM B. M. (1991). Third Message Understanding Evaluation and Conference (MUC-3) : Phase 1 Status Report. In *Proceedings of the 4th DARPA Workshop on Speech and Natural Language*, p. 301–305, San Diego, California, USA.

WAYNE C. (1998). Topic Detection & Tracking : A Case Study in Corpus Creation & Evaluation Methodologies. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.