Un modèle probabiliste de détection en ligne de nouvel événement

A probabilistic model for online New Event Detection

H. Binsztok[1] T. Artières[1] P. Gallinari[1]

[1] LIP6
8, rue du Capitaine Scott
Paris, France
{Henri.Binsztok,Thierry.Artieres,Patrick.Gallinari}@lip6.fr

Résumé

Cet article traite de la détection de nouvel événement (NED), une sous tâche du programme de détection et suivi d'événements (TDT). Par exemple, la NED vise à détecter l'apparition d'un nouvel événement dans un flux de documents textuels d'agence de presse. Dans ce but, nous cherchons à maintenir des modèles de chacun des événements précédemment identifiés, évoluant au cours du temps. Chaque fois qu'un nouveau document apparaît, nous devons établir s'il traite d'un événement déjà connu ou d'un nouvel événement. Nous présentons ici un nouveau modèle probabiliste pour la modélisation des événements. Ce modèle constitue la base sur laquelle nous introduisons notamment une mesure d'activité des événements. Nous montrons l'efficacité de notre approche empiriquement sur des dépêches en français issues de l'actualité internationale en la comparant aux approches de l'état de l'art.

Mots Clef

Recherche d'Information - Détection et Suivi d'Evénements - Détection de Nouveauté

Abstract

This article presents a novel approach for New Event Detection (NED), a subtask of the Topic & Detection Tracking (TDT) program. NED applies to detecting new events in press news for instance. For such a task, we have to keep up-to-date models of events, supposing each document is linked to a single event. By definition, the event is a drifting concept and NED requires to continually determine whether a new document is an instance of an existing event or a new concept that is added to a set of active events. We build a probabilistic framework derived from the Naïve Bayes formalism, as the ground for the introduction of the activity of an event, which brings temporal information beside the classical time window. We performed tests on French press agency news to show the validity of our proposal compared to different classical models.

Keywords

Information Retrieval - Topic Detection & Tracking - New Event Detection

1 Introduction

De nombreuses tâches d'intelligence artificielle ou de reconnaissance des formes traitent de la classification de données. Souvent, les classes sont connues et une partie des exemples étiquetés : c'est le cadre classique de la classification supervisée de documents. Parfois, on ne dispose pas de la liste des classes : c'est le cadre du clustering de documents. La tâche de détection de nouveauté est une tâche de clustering, qui ne nécessite aucune connaissance *a priori* du nombre de classes. Plus spécifiquement, on considère le cas d'une source qui émet des documents *datés*, chaque document étant lié à un *événement*. On cherche à détecter automatiquement l'événement lié à un document donné, et plus spécifiquement à déterminer si un document traite d'un événement déjà connu ou d'un nouvel événement.

A titre d'exemple, considérons le flux de documents émis par une agence de presse. Ainsi, la dépêche intitulée "Un sommet UE-USA sur fond de tensions transatlantiques" est liée à l'événement "Tensions entre l'Europe et les Etats-Unis". Ce label d'événement ne sera pas déterminé par le système : le but est de grouper les différents document en clusters, dont le nombre évolue au cours du temps. Chaque fois qu'un nouveau document est émis, on cherche à savoir s'il est lié à un événement déjà connu ou un nouvel événement. Cette tâche a été définie sous l'appellation NED - New Event Detection - par un programme traitant plus généralement de la détection et du suivi d'événement[1]. Ce programme a été lancé par la DARPA en 1996 [Wayne, 1998]. Cet article traitera de la détection de nouvel événement, même si les modèles mis en oeuvre peuvent être directement adaptés au suivi de documents - c'est-à-dire leur classification sur l'ensemble des événements connus.

[1] Topic Detection and Tracking - TDT

La détection peut être hors-ligne ou en ligne. Dans ce dernier cas, nous devons tenir compte de deux contraintes : d'une part on ne connaît aucun document postérieur au document considéré, et d'autre part, la détection devant être appliquée chaque fois qu'un nouveau document est observée, la complexité devra être réduite autant que possible. Ces deux modes de fonctionnement supposent deux applications différentes : dans un cas, on cherche à déterminer l'ensemble des événements et en particulier leur nombre, par exemple pour présenter une synthèse d'information, dans l'autre, on cherche à avertir un opérateur d'un nouvel événement susceptible de l'intéresser. Nous avons choisi plus spécifiquement de considérer la problématique *en ligne*.

Notre objectif est de proposer une approche stable et efficace pour la NED, via l'introduction d'un modèle probabiliste intégrant une sélection de caractéristiques et une mesure temporelle de l'information. Son coût calculatoire faible la destine à une application en ligne.

Nous commençons par une description de notre modèle probabiliste de documents (section 3), qui modélise spécialement la détection de nouveauté. Ensuite, nous proposons un algorithme probabiliste de clustering (section 4) qui permet de détecter si un document traite d'un événement nouveau, ou non. La section 5 est dédiée à la sélection de caractéristiques, qui permet à la fois de simplifier le modèle et d'adapter sa complexité en fonction des données. La section 6 présente l'intégration de la mesure de l'activité d'un événement introduite dans [Binsztok and Gallinari, 2002]. Enfin, les résultats expérimentaux seront présentés et discutés dans la section 7.

2 Etat de l'art

Plusieurs travaux ont déjà traité de la problématique de NED dans le cadre du TDT : [Spitters and Kraaij, 2001], [Walls et al., 1999], [Yang et al., 1999], [Yang et al., 2000]. Mais, d'autres auteurs se sont intéressés à cette problématique : on peut citer [Parra et al., 1996] qui propose une méthode de détection de nouveauté appliquée à la prédiction de pannes de moteurs. Une synthèse pertinente des différentes approches utilisées a été proposée par [Allan et al., 1998]. Désormais, le travail s'est focalisé autour d'approches numériques.

On peut identifier deux composantes dans un système TDT : la première est une représentation pour les documents et les événements ; la seconde un algorithme principal pour la détection reposant sur la représentation précédente. Certaines approches comme [Stokes and Carthy, 2001] utilisent une représentation incluant des techniques de traitement automatisé du langage naturel fondée sur le chaînage lexical. Mais la plupart des approches utilisent le modèle vectoriel [Salton, 1968]. Dans cette représentation, chaque document ou événement est perçu comme un vecteur de caractéristiques ayant pour dimension la taille du vocabulaire - chaque composante représentant le poids statistique affecté à un mot donné. Souvent, les poids dans cet espace de représentation sont déterminés par un codage tf-idf, ou ses variantes comme Okapi [Robertson et al., 1995]. La similarité entre documents est alors déterminée par un cosinus.

Parmi les algorithmes utilisés pour la NED, [Yang et al., 1999] utilise une variante du Group Average Clustering ([Cutting et al., 1992]). Cette approche est adaptée à l'application hors-ligne car elle profite de la connaissance de l'ensemble des documents et sa complexité est élevée. [Allan et al., 1998] propose de détecter des changements abrupts de vocabulaires ; même si cette approche requiert un court délai, nous verrons que notre formalisme reprend cette notion dans la modélisation du nouvel événement, mais sans délai. [Yang et al., 1998] et [Yang et al., 1999] fournissent un exemple représentatif d'algorithme de clustering, utilisant le modèle vectoriel pour représenter les documents. Les documents sont évalués séquentiellement par ordre d'arrivée, et les clusters correspondant aux événements croissent incrémentalement : un nouveau document est ajouté au cluster le plus proche si une mesure de similarité entre un document et un cluster est supérieure à un seuil, sinon un nouveau cluster est créé. Dans le cas le plus simple, un cluster correspond à un événement, mais un événement peut aussi correspondre à plusieurs clusters. Un second seuil permet de déterminer si un nouveau cluster reste lié à un événement existant ou constitue un nouvel événement. Comme indiqué dans un rapport TDT [2] toujours d'actualité, la tâche de NED reste à améliorer, au niveau des performances et au niveau du formalisme. La principale difficulté provient de la nécessité d'apprendre de façon non-supervisée un modèle d'événement évoluant dans le temps, et capable de s'adapter à tout type de données. Dans le cas de la problématique en ligne, s'y ajoute un aspect calculatoire : le système devra être assez rapide pour décider sitôt un nouveau document arrivé.

L'utilisation d'un formalisme probabiliste n'est pas nouveau. Plusieurs auteurs réunis dans un atelier spécial l'ont introduit dans [Allan et al., 1999]. Ils concluent qu'un modèle probabiliste permet de justifier les approches courantes, tout en restant légèrement en retrait en terme de performance. L'autre modèle probabiliste dont nous avons connaissance est celui de [Baker et al., 1999]. Les auteurs utilisent un modèle multinomial pour apprendre une hiérarchie d'événements. Les méthodes employées - shrinkage et recuit simulé - sont théoriquement très intéressantes mais leur apprentissage est difficile et complexe algorithmiquement. Enfin, la hiérarchisation a pris une place importante dans les derniers travaux du domaine. [Yang et al., 2002] qui rappelle que la tâche est un challenge ouvert, propose une nouvelle approche qui vise à grouper les événements en catégories *via* un apprentis-

[2] `http://www.nist.gov/speech/tests/tdt/tdt2001/PaperPres/CMU-pres/TDTwkshop2001.ppt`

sage supervisé, puis d'appliquer la détection de nouveauté au sein d'une catégorie. Pourtant, la nécessité de disposer d'un étiquetage supplémentaire des documents déplace légèrement la problématique.

3 Formalisme probabiliste

Cette section présente notre formalisme probabiliste pour la tâche de détection. Le point de départ est le travail initié par [Allan et al., 1999] que nous allons étendre. On considère une source qui émet des documents datés ; à l'instant t, $n(t)$ documents ont déjà été vus et chacun d'eux a été assigné à un des $m(t)$ événements détectés. A l'instant $t+\epsilon$, un nouveau document $d_{n(t)+1} = d$ est émis : le système doit décider si d est lié à un événement $e_j, j \in [1, m(t)]$ - et alors décider lequel - ou à un nouvel événement, qu'il faut alors créer. A chaque instant, on se place alors dans le cadre d'un problème de classification, dans lequel les événements jouent le rôle de classes. Pour classifier un document dans une classe, on considère un ensemble de modèles d'événements qui soient des générateurs stochastiques de documents, ainsi déterminés par une distribution de probabilité $P(d|e)$, soit la probabilité d'un document sachant son modèle d'événement. L'événement "nouveau" est considéré comme un événement supplémentaire associé à la distribution $P(d|\mathrm{new})$. Alors, construire le système revient à construire une règle de décision, soit la fonction d'assignation $\mathrm{event}(d)$ qui affecte à chaque document un événement :

$$\mathrm{event}(d) = \mathrm{argmax}_{e,\mathrm{new}} P(d, e) \tag{1}$$

Ainsi, pour implémenter cette règle de décision, nous devons calculer deux types de probabilités : $P(d|e)$ et $P(e)$ pour tous les événements, dont l'événement spécial "nouveau". Le reste de cette section traite de l'estimation de ces quantités, que nous utiliserons dans les algorithmes de clustering décrits dans la section 4.

3.1 Probabilités *a priori* des événements

Au premier abord, la probabilité *a priori* d'un événement ordinaire - tous sauf l'événement new - peut être estimée via la proportion de documents qui sont associés à cet événement [Allan et al., 1999] : $P(e) = \frac{N(e)}{N(E)}$ où $N(e)$ et $N(E)$ représentent respectivement le nombre de documents affectés à l'événement e et le nombre total de documents. Cependant, pour détecter la nouveauté, il faut ajouter un nouvel événement, dont la probabilité $P(\mathrm{new})$sera présentée ultérieurement. Ainsi, $\sum_{e \in \{E \cup \mathrm{new}\}} P(e) = 1$ Pour l'instant, étant donné $P(\mathrm{new})$, on considère que la probabilité *a priori* d'un événement est donnée par :

$$\begin{aligned} P(e) &= P(\text{Not new })P(e|\text{Not new }) \\ &= (1 - P(\mathrm{new}))\frac{N(e)}{N(E)} \end{aligned} \tag{2}$$

Nous avons deux possibilités concernant $P(\mathrm{new})$: soit nous estimons $P(\mathrm{new})$directement, soit nous permettons qu'elle soit fixée arbitrairement, devenant un hyperparamètre du système. Tout en conservant au système sa simplicité, jouer sur $P(\mathrm{new})$permet d'encourager ou au contraire de décourager l'apparition de nouveaux événements. Ainsi, ce paramètre sera utilisé pour contrôler le fonctionnement du système.

3.2 Modèle d'événement

Nous traitons ici de l'estimation de $P(e|d)$ pour un événement e existant. Nous considérerons le cas d'un nouvel événement dans la section 3.3. D'après la règle de Bayes,

$$P(e|d)P(d) = P(d|e)P(e) \tag{3}$$

Chaque document comporte trois types d'information : son contenu - les différents mots qu'il emploie, sa séquence - l'ordre des mots - et sa structure - son organisation en paragraphes, sections, etc. Nous allons utiliser l'hypothèse Naïve-Bayes, qui ne retient que le contenu[3] : l'ensemble des mots w qui apparaissent dans e :

$$P(d|e) = \prod_{w \in d} P(w|e) \tag{4}$$

Estimer la probabilité conditionnelle $P(w|e)$ repose sur le choix d'un modèle de langage. Nous avons conservé le modèle de langage de Witten-Bell déjà utilisé dans [Allan et al., 1999] et décrit dans [Manning and Schütze, 1999]. L'estimation suivante est alors utilisée :

$$P(w|e) = \frac{1}{\mathrm{card}\{V_{|e}\} + l(e)} \left(\mathrm{tf}(w, e) + \mathrm{card}\{V_{|e}\}P(w)\right) \tag{5}$$

où $\mathrm{tf}(w, e)$ est le nombre d'occurrences du mot w dans l'événement e, $l(e)$ la longueur de l'événement e - soit le nombre total de mots de tous les documents affectés à l'événement e - et $\mathrm{card}\{V_{|e}\}$ le nombre de mots distincts employés dans les documents affectés à e. Avec cette estimation, $P(w)$ est approximé en utilisant le laplacéen ([Manning and Schütze, 1999]) :

$$P(w) = \frac{tf(w, D) + 1}{\mathrm{card}\{V\} + L} \tag{6}$$

où $\mathrm{card}\{V\}$ est la taille du vocabulaire et L la longueur totale de tous les documents actifs (suffisamment récent). $\mathrm{tf}_t(w, D) = \mathrm{tf}_t(w, E)$ est le nombre total d'occurrences du mot w dans tous les documents actifs, ou dans l'ensemble des événements. Les valeurs sont identiques à l'instant t, elles ne le sont plus à $t + \varepsilon$. En effet, le nouveau document qui vient d'être émis est compté dans D, mais pas dans E car on n'a pas encore déterminé son événement. C'est ce décalage qui va être utilisé pour estimer les probabilités conditionnelles de l'événement nouveau.

[3] D'où son appellation de "sac de mots"

3.3 Probabilité de l'événement nouveau

Pour pouvoir estimer la probabilité qu'un nouveau document traite d'un autre événement que ceux déjà connus, nous devons calculer $P(d \text{ is new}|d)$. En appliquant une nouvelle fois la règle de Bayes :

$$P(d \text{ is new}|d) = \frac{P(d|d \text{ is new})P(d \text{ is new})}{P(d)} \quad (7)$$

où $P(d)$ est calculé comme $P(d) = \sum_{x \in \{E(t) \cup \text{new}\}} P(d|x)P(x)$. En appliquant la même hypothèse Naïve-Bayes que précédemment, on a $P(d|d \text{ is new}) = \Pi_{w \in d} P(w|d \text{ is new})$. On pourrait alors estimer les probabilités $P(w|d \text{ is new})$ comme on l'a déjà fait pour un événement ordinaire dans la section précédente, en considérant que l'événement nouveau est représenté par un seul document, celui qui vient d'être émis. Cependant, cela mène à un problème d'estimation car il y a très peu d'information pour estimer les vraisemblances des mots avec l'événement new, ce qui conduit à des performances expérimentales en retrait. Nous proposons une autre façon de calculer ces vraisemblances. On peut exprimer la probabilité d'un mot dans l'événement new par

$$P(w|d \text{ is new}) = \frac{P(d \text{ is new}|w)P(w)}{P(d \text{ is new})} \quad (8)$$

Pour calculer cela, il nous faut estimer la probabilité a posteriori de l'événement new conditionné par le mot w dans le document. Nous proposons :

$$P(d \text{ is new}|w) = \frac{\text{tf}(w,d)+1}{\text{tf}(w,D)+1} \quad (9)$$

où $\text{tf}(w,d)$ représente le nombre d'occurrences de w dans le document d et $\text{tf}(w,D)$ le nombre d'apparitions de w dans tous les documents connus à l'instant $t+\varepsilon$. L'équation 9 présente des propriétés intéressantes : la probabilité est égale à 1 pour un mot qui serait présent dans le document d, mais qui ne serait jamais apparu dans les documents précédents. Elle tend vers 0 pour un mot très courant dans les documents précédents. L'intégration d'un lissage permet de simplifier une partie des termes et on obtient

$$P(d \text{ is new}|d)P(d) = \frac{1}{P(d \text{ is new})^{l(d)-1}} \cdot \prod_{w \in d} \frac{\text{tf}(w,d)+1}{\text{card}\{V\}+L} \quad (10)$$

où $l(d)$ représente la longueur du document d. Il est intéressant de noter que ce modèle fait automatiquement apparaître une normalisation sur la longueur du document.

4 Algorithmes de clustering

Dans cette section, nous présentons l'algorithme de clustering qui permet de détecter si un document traite, ou non, d'un événement nouveau. Notre choix a été motivé par la simplicité, car des algorithmes trop complexes ne sont pas adaptés à la problématique en ligne. Nous commençons par la description de l'algorithme de base [Yang et al., 1999] formalisé dans notre cadre probabiliste. Ensuite, nous présentons notre approche, qui s'affranchit du seuil de nouveauté. Tous les modèles et comptages doivent être mis à jour dès qu'un nouveau document est émis.

4.1 Algorithme avec seuil de nouveauté

L'algorithme classique de clustering employé pour la détection de nouveauté fait intervenir un seuil de nouveauté t_n, et le schéma de l'algorithme dans les notations de notre formalisme probabiliste est le suivant lors de l'émission d'un nouveau document d :

1. Calculer $P(d \text{ is new}|d)$
2. Si $P(d \text{ is new}|d) > t_n$, le document est considéré comme traitant d'un événement nouveau. Cet événement est crée et d devient le premier document lié à cet événement.
3. Sinon ($P(d \text{ is new}|d) \leq t_n$)
   (a) Calculer $e^* = \text{argmax}_{e \in E(t)} P(e|d)$
   (b) Ajouter d à l'événement e^*.

4.2 Algorithme sans seuil de nouveauté

Pour éviter l'utilisation d'un seuil de nouveauté devant être ajusté expérimentalement et redondant avec la probabilité *a priori* de nouveauté, nous proposons un algorithme alternatif dans lequel les probabilités conditionnelles sont comparées directement :

1. Calculer $P(x|d)$ pour chaque $x \in E \cup \text{new}$,
2. Si $x^* = \text{argmax}_{x \in E \cup \text{new}} P(x|d) = \text{new}$, le document est considéré comme traitant d'un événement nouveau.
3. Sinon ($x^* = \text{argmax}_{x \in E \cup \text{new}} P(x|d) \neq \text{new}$)
   (a) Ajouter d à l'événement x^*.

5 Sélection de caractéristiques

Lors du calcul de l'équation 4, les vraisemblances sont calculées comme un produit[4] sur tous les mots du vocabulaire. Or, le vocabulaire ayant tendance à croître au fur et à mesure que des nouveaux documents arrivent, deux problèmes apparaissent. Le premier est que certains mots sont faiblement représentés et qu'ainsi les probabilités associées $P(w|e)$ sont mal estimées, résultant en des modèles d'événements imprécis. Le second est la hausse de la complexité liée. Pour cela, nous allons proposer un algorithme de sélection de caractéristiques, appliqué aux mots représentant les événements. L'idée est proche de celle du critère de description minimale (MDL, [Rissanen, 1982]) : il s'agit d'accepter de diminuer la vraisemblance des documents avec le modèle d'événement pour diminuer la

[4] En pratique, on calcule des sommes de logarithmes

taille du modèle. Ainsi, l'apprentissage est plus facile - le modèle est moins sensible au bruit. On pourra se reporter à [Binsztok and Gallinari, 2002] pour une présentation à l'utilisateur de l'évolution des modèles d'événements au cours du temps.

5.1 Une sélection basée sur l'information mutuelle

De nombreuses approches ont été proposées pour la sélection de caractéristiques. Nos contraintes de complexité ont éliminé des méthodes basées sur les décompositions de Karhunen - Loeve ou d'analyse discriminante présentées dans [Kittler, 1974]. Les approches plus simples reposent sur des principes de théorie de l'information [Cover and Thomas, 1991]. Suivant le travail de [McCallum and Nigam, 1998], nous avons considéré les mots indépendants : la présence ou l'absence d'un mot dans les caractéristiques retenues ne dépend pas de la présence d'autres mots - contrairement à des méthodes utilisant des cooccurrences. Pour un mot w, on cherche à établir sa pertinence $R(w)$ pour la classification d'un document dans un événement donné via l'information mutuelle entre deux variables aléatoires. Soit V_w la variable associée à la présence du mot w dans le nouveau document d - qui prend soit la valeur $vrai$ soit la valeur $faux$ - et V_e l'événement auquel est rattaché le document :

$$I(V_e, V_w) = \sum_{v_e, v_w} (P(V_e = v_e, V_w = v_w) \log \frac{P(V_e = v_e, V_w = v_w)}{P(V_e = v_e)P(V_w = v_w)}) \quad (11)$$

Alors, en notant $P(\bar{w}) = 1 - P(w)$:

$$R(w) = \sum_{e \in E} (P(e, w) \log \frac{P(e, w)}{P(e)P(w)} + P(e, \bar{w}) \log \frac{P(e, \bar{w})}{P(e)P(\bar{w})}) \quad (12)$$

Puisque les mots sont considérés indépendants, la recherche du sous-ensemble du vocabulaire décrivant le mieux un événement est obtenu en sélectionnant itérativement les mots qui maximisent le critère de l'équation 11. Pour terminer la définition de notre méthode de sélection de mots, il reste à définir un critère d'arrêt. Nous avons choisi de prendre un nombre fixé de mots, cette valeur étant optimisée sur un ensemble de données d'entraînement, distinctes du corpus de test.

5.2 Sélection par événement

Il est possible d'étendre la sélection de caractéristiques précédente en effectuant une sélection différente pour chaque événement. Notre but est donc de trouver pour chaque événement e les r mots qui le décrivent le mieux par rapport aux autres ; r est supposé pour l'instant constant pour l'ensemble des événements. Pour cela, on cherche à sélectionner les r mots qui optimisent un critère dérivé du précédent :

$$R_e(w) = P(e, w) \log \frac{P(e, w)}{P(e)P(w)} + P(e, \bar{w}) \log \frac{P(e, \bar{w})}{P(e)P(\bar{w})} \quad (13)$$

Enfin, pour répondre particulièrement à nos attentes sur la complexité, il est possible d'appliquer de façon *paresseuse* cette sélection de caractéristiques. Pour cela, un modèle d'événement est conservé tant qu'aucun nouveau document ne lui est ajouté. Cette stratégie, bien que non optimale, fournit des résultats expérimentaux comparables. Dans la section suivante, nous allons examiner le rôle du temps dans la tâche et notamment comment utiliser l'information temporelle pour la sélection de caractéristiques.

6 Le rôle du temps

L'information temporelle - les dates des documents - constitue une des informations majeures, et contribue à la spécificité de la tâche. Son utilisation dans les systèmes de NED reste assez brute : généralement, elle n'est considérée que par l'utilisation d'une fenêtre temporelle. A l'instant t, la fenêtre d'une longueur fixée T couvre seulement une partie des documents précédents : ceux émis entre $t - T$ et t. Les documents présents dans la fenêtre sont pris en compte pour l'estimation des modèles d'événements : ils sont dits *actifs*. Les autres sont ignorés : en pratique, cela permet de limiter la taille du vocabulaire et donc de diminuer la complexité du système. Nous avons repris cette approche, mais nous proposons également une mesure d'activité d'événement, introduite dans [Binsztok and Gallinari, 2002].

L'activité d'un événement est élevée quand beaucoup de documents récents traitent de cet événement ; nous rappellerons sa définition dans 6.1. Intuitivement, cette mesure est fortement liée à la probabilité *a priori* des événements. Nous verrons dans la section 6.2 une approche où les mesures d'activité servent à estimer les probabilités *a priori* des événements. Nous verrons qu'il est possible de fournir une autre interprétation de l'activité : plus l'activité est faible, moins il y a de documents pour estimer les lois de probabilités du modèle et plus cet estimation est délicate. Ainsi, nous pouvons utiliser cette mesure d'activité pour adapter la constante de sélection de caractéristiques r décrite section 5.2. Ainsi, plus un événement est actif, plus le nombre de mots retenus pour le modéliser sera important.

6.1 Mesure de l'activité d'un événement

Le cycle de vie d'un événement typique possède trois étapes : une naissance, quand le premier document associé apparaît, une croissance, quand une quantité croissante de documents lui sont associés puis une mort - progressive ou soudaine, lorsque de nouveaux événements prennent le relais de l'actualité. Ainsi, chaque événement a un cycle de vie au cours duquel son activité varie. Pour déterminer cette activité, nous faisons l'hypothèse que chaque docu-

ment possède une période active durant laquelle il contribue à l'activité d'un événement. Soit τ ce délai ; à l'instant t, l'activité de l'événement e est

$$A(e,t) = \int_{t-\tau}^{t} f(s)\delta(e,s)\mathrm{d}s, \qquad (14)$$

où $\delta(e,s)$ est le nombre de documents[5] liés à l'événement e à l'instant s. Nous considérerons le cas $\tau < T$, où T est la taille de la fenêtre temporelle. Enfin, f est une fonction linéaire de pondération, qui vise à privilégier les documents les plus récents, telle que $f(t-\tau) = 0$ et $f(t) = 1$. Nous allons aborder l'utilisation de l'activité pour la détection de nouveauté.

6.2 Modification des probabilités *a priori*

Dans le formalisme décrit dans la section 3, nous avons considéré des probabilités *a priori* pour les événements indépendantes du temps et estimées en fonction de la proportion de documents traitant de chacun des événements. Nous allons voir qu'il est possible d'étendre cette estimation. Notamment, certains événements sont périodiques - hebdomadaires, mensuels, etc. - et une analyse fréquentielle permettrait de les détecter. Nous avons choisi une approche plus simple et plus robuste, qui vise à estimer les probabilités *a priori* en fonction des activités des événements :$P(e) = \frac{A(e,t)}{\sum_{e' \in E} A(e',t)}$. Cette estimation revient à attribuer un poids statistique à chacun des documents, qui varie en fonction du temps. Nous allons maintenant proposer une utilisation alternative de l'activité des événements.

6.3 Taille des modèles d'événements

Lors de l'introduction de la sélection de caractéristiques pour les modèles d'événements en section 5.2, nous avons proposé de sélectionner un nombre constant de r caractéristiques pour chacun des modèles d'événements. Cependant, comme indiqué dans l'introduction de cette section, plus un événement génère de documents - plus son activité est élevée - et plus de détails sont donnés, permettant de mieux estimer les probabilités conditionnelles. Ainsi, nous avons proposé d'adapter la taille du modèle à la complexité des données. Pour cela, pour chaque événement e, nous sélectionnons un nombre variable de caractéristiques, fonction de l'activité de son activité à l'instant t :

$$r(e) = r_c + r_v A(e,t), \qquad (15)$$

où r_c et r_v représentent respectivement les nombres constant et variable de mots sélectionnés. Ainsi, les événements les plus actifs seront représentés avec plus de détails que les événements moins fréquents.

[5] 0 ou 1 puisque on consid`ere que deux documents ne peuvent pas ˆetre ´emis au mˆeme instant

7 Résultats

7.1 Mesures d'évaluation

Taux d'erreur et fausses alarmes Le système de détection renvoie une valeur booléenne {vrai ; faux} pour un couple $\{d_i, e(d_i)\}$, assorti d'une mesure de confiance. On a alors les mesures suivantes :

	EST positif	EST négatif
JUGÉ positif	a	b
JUGÉ négatif	c	d

L'évaluation repose alors sur deux indicateurs : le taux d'erreur $m = \frac{c}{a+c}$ (si $a + c > 0$, indéfini sinon) et le taux de fausse alarme $f = \frac{b}{b+d}$ (si $b + d > 0$, indéfini sinon) auxquels peuvent s'adjoindre des estimateurs plus classiques comme la précision $p = \frac{a}{a+b}$, et le rappel $r = \frac{a}{a+c}$. Il est possible d'agréger précision et rappel avec la mesure F, moyenne harmonique du rappel et de la précision : $F = \frac{2rp}{r+p}$.

La détection se faisant exclusivement sur le premier article d'un événement, on dispose de peu d'exemples positifs. A cet effet, plusieurs "versions" successives du corpus sont utilisées pour le test. La version $n+1$ est obtenue en éliminant de la version n la première nouvelle relative à un événement. On détectera donc ainsi la deuxième, puis la troisième, etc. nouvelle relative à un événement. Cette procédure permet d'avoir plus d'exemples pour mieux évaluer les performances d'une méthode, les résultats étant ensuite moyennés.

En faisant varier la probabilité *a priori* de nouveauté $P(\mathrm{new})$, on peut donc construire une courbe DET - taux d'erreur fonction du taux de fausses alarmes. Cet hyperparamètre du système contrôle la sensibilité de l'algorithme. S'il est augmenté, le système est forcé de détecter plus d'événements nouveaux, au prix d'un taux de fausses alarmes plus élevé. Au contraire, s'il est diminué, le système détectera moins d'événements nouveaux, mais générera parallèlement moins de fausses alarmes.

Suivant l'application envisagée, on choisira un point de fonctionnement. Si la vocation du système est informative, il est préférable de se placer à un faible taux d'alarmes : ainsi, le système peut négliger certains événements mais fournit une information tout de même pertinente. A l'opposé, pour un système décisionnel, le système se présente comme une assistance à un utilisateur. Dans ce cas, il est préférable de minimiser le taux d'erreurs ; et ce, d'autant plus qu'il est probable que, parmi les fausses alarmes, un grand nombre correspondant à un changement significatif d'un événement existant. D'une façon générale, en maintenant les deux taux inférieurs à 20%, le système est tout à fait utilisable dans le cadre d'un usage courant appliqué à des nouvelles d'actualités.

Evaluation de la détection seule La méthode d'évaluation proposée précédemment lie fortement les tâches de détection et de suivi. Si le système devait diverger - à cause d'un suivi peu performant, les détections futures seraient compromises car les modèles d'événements deviendraient très mal estimés. Cela nous amène à modifier légèrement la méthode de détection : pour cela, pour chaque nouveau document, on évalue la réponse du système concernant un événement nouveau, ou non. Ensuite, le document est affecté à son événement *réel*. Ainsi, à chaque instant, les ensembles d'événements constitués sont corrects. De plus, la comparaison des deux méthodes d'évaluation permet de mieux comprendre le fonctionnement de notre approche sur des données réelles.

7.2 Données disponibles

Nous avons constitué pour ces travaux un corpus (nommé `monde2`) créé à partir de Yahoo !News[6]. Il s'agit, comme pour les corpus du programme américain TDT, de dépêches d'agence, mais en langue française. Il totalise 4819 dépêches d'agence issues de l'actualité internationale correspondant à une trentaine de thématiques et 113 événements. Cette connaissance sert de validation, mais n'est pas fournie au système. Chaque document contient les champs suivants : un titre, une date, l'événement dont traite le document, et un corps de texte. Nous avons prétraité les documents à l'aide d'un antidictionnaire français ainsi qu'un préprocesseur développé autour de treetagger[7], qui remplace chaque mot connu par son lemme tout en conservant les mots inconnus - qui représentent souvent des noms propres. A titre d'exemple, nous donnons dans le tableau 1 une partie des événements du corpus. Pour illustrer la notion d'activité présentée dans notre approche, la figure 1 montre le profil temporel de deux événements différents : un événement récurrent, qui apparaît régulièrement et dont le modèle est susceptible d'évoluer au cours du temps ; et un événement unique, apparaissant soudainement et disparaissant rapidement.

7.3 Expériences préliminaires

Les premières expériences ont cherché à valider notre modèle probabiliste, et à comparer les deux stratégies de détection détaillées dans la section 4. La figure 4 compare les performances de notre modèle à celles du système de référence de [Yang et al., 1999], qui repose sur une mesure d'Okapi dans le modèle vectoriel. Pour chacun des systèmes, nous avons utilisé une fenêtre booléenne de 30 jours - c'est-à-dire qu'un document est retiré des comptages de son événement au bout de 30 jours, un événement étant supprimé dès lors qu'il ne contient plus aucun document dans la fenêtre. Même si l'algorithme 4.1 peut parvenir à de meilleurs résultats au prix d'un travail d'ajustement qui constitue une sur-adaptation aux données, nous avons

[6] `http://fr.news.yahoo.com`

[7] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

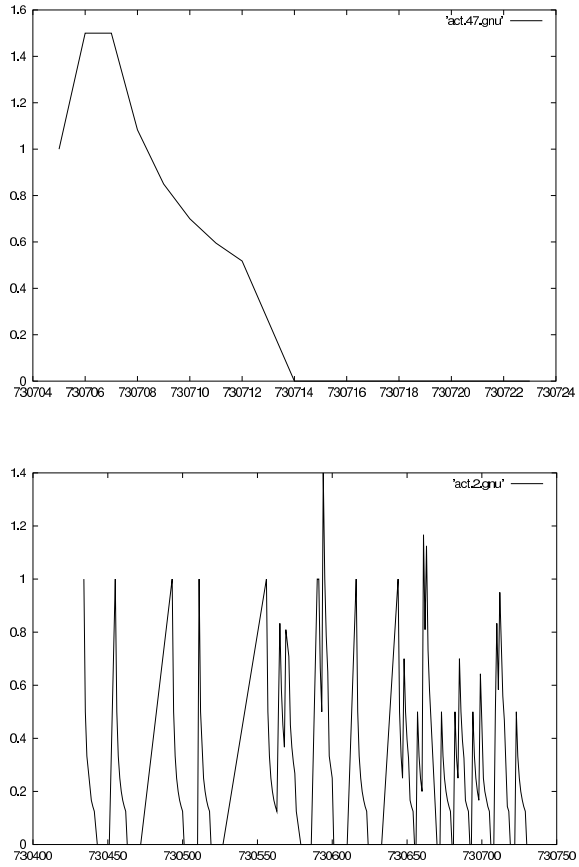

FIG. 1 – Profil de l'activité d'un événement unique (en haut) et d'un événement récurrent (en bas). L'activité est en ordonnée, le numéro du jour en abscisse

retenu la seconde stratégie 4.2 qui fournit des résultats similaires sans nécessiter de régler le seuil de nouveauté. En effet, sans évaluation sur les données, il est délicat de régler conjointement ce seuil et la probabilité *a priori* de nouveauté. La figure 2 montre les effets de la fenêtre sur le nombre d'événements à un instant donné ; la figure 3 rappelant que l'effacement d'événement affecte également les performances.

7.4 Modèles d'événements

Dans cette partie, nous allons évaluer les méthodes de sélection de caractéristiques présentées en section 5 ainsi que d'intégration du temps présentée en section 6. La figure 5 compare trois systèmes : le premier - complet - n'utilise aucune sélection de variables. Le second - constant r = 11 - utilise une sélection de variables indépendante du temps, où chaque événement est représenté par seulement $r = 11$ mots. Le troisième - variable rc = 8 rv = 2 - utilise une sélection de caractéristiques dépendante du temps, selon l'équation 14 qui offre des performances à peine

Accident de Kitzsteinhorn
Attentat contre le MI6 à Londres
Collision ferroviaire en Angleterre
Débat sur le bouclier antimissile
Détournement d'un avion russe à Istanbul
Élections fédérales au Canada
Élections générales en Roumanie
Élections législatives en Thaïlande
Élections présidentielles au Ghana
Élections présidentielles au Sénégal
Élections présidentielles au Venezuela
Élections présidentielles aux États-Unis
Géorgie : la disparition de 3 membres du CICR

TAB. 1 – Exemples d'événements de notre corpus, dont plusieurs représentatifs de la thématique "élections"

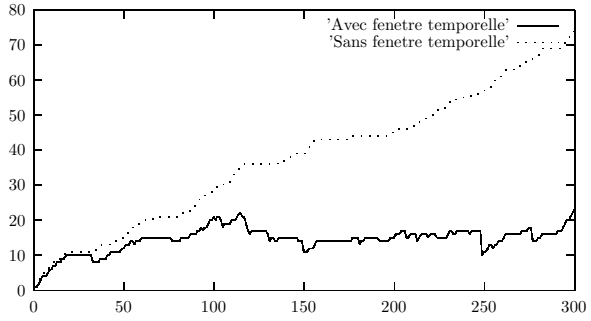

FIG. 2 – Nombre d'événements actifs avec et sans fenêtre temporelle ; le nombre de jours est en abscisse

supérieures[8]. Pour le modèle variable, nous avons fixé la taille de la fenêtre d'activité τ égale à la fenêtre temporelle T, comme le montre la figure 6. D'une façon générale, le système avec sélection de variables offre des performances du même ordre que le système complet, tout en apportant deux avantages : le premier en terme de complexité, surtout avec l'évaluation paresseuse des modèles utilisée dans ces expériences. Le second, la possibilité de présenter visuellement à l'utilisateur le modèle d'événement et son évolution au cours du temps [Binsztok and Gallinari, 2002]. Il est intéressant de noter par ailleurs que les valeurs de r ont été choisies en optimisant sur un nombre très faible de documents (100). Enfin, l'amélioration apportée par la sélection de caractéristiques est particulièrement sensible dans le domaine des faibles taux d'erreur. Ce domaine de fonctionnement est particulièrement utile lorsque le système de NED est utilisé pour assister un utilisateur effectuant un travail de veille technologique. En effet, dans ce cas les fausses alarmes sont acceptables car l'opérateur peut les vérifier, tandis que s'il se fie au système, il manquera les documents

[8]Ce mod`ele conserve un avantage signifi catif pour l'interface avec l'utilisateur

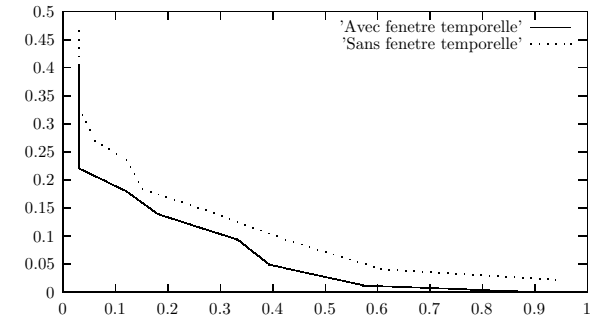

FIG. 3 – Performance avec et sans fenêtre temporelle : le taux d'erreurs est en abscisse, le taux de fausses alarmes en ordonnées

qui ne lui sont pas présentés.

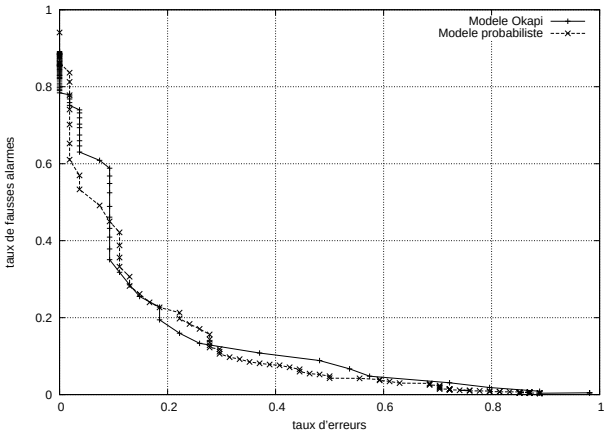

FIG. 4 – Modèle probabiliste et modèle vectoriel avec okapi

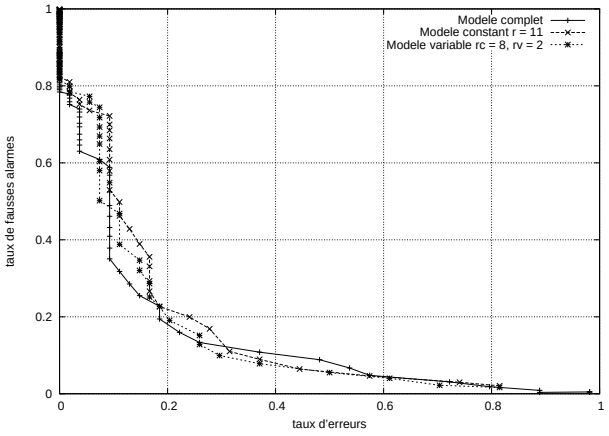

FIG. 5 – Modèles d'événements : complet, constant $r = 11$ et variable $r_c = 8, r_v = 2$

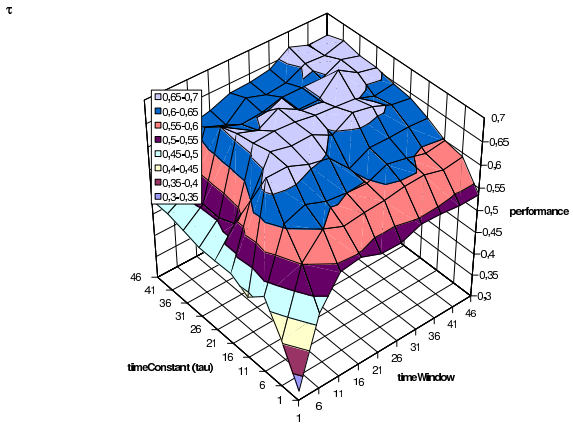

FIG. 6 – Optimisation de la valeur de la fenêtre d'activité τ en fonction de la fenêtre temporelle T : T et τ sont exprimées en jours, la performance en mesure F normalisée

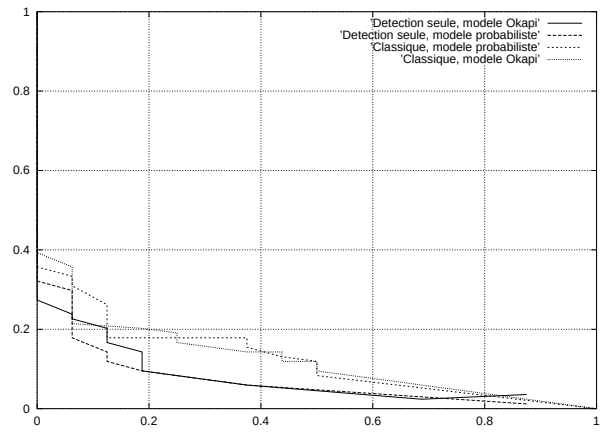

FIG. 7 – Evaluation classique ou de la seule détection : le taux de fausses alarmes est en abscisse, le taux d'erreurs en ordonnées

La figure 7 compare les deux méthodes d'évaluation proposées dans 7.1. Tant que la fenêtre temporelle est assez limitée, les deux méthodes sont équivalentes. Par contre, plus la fenêtre temporelle augmente, plus la dispersion des clusters gêne les détections futures, et des différences surviennent entre les méthodes d'évaluation. Le choix de l'une ou de l'autre dépend fortement du type d'application du système. Par exemple, si le but est d'alerter un utilisateur de l'apparition de nouveaux documents susceptibles de l'intéresser, il est possible que l'interface accepte un retour de l'utilisateur et l'intègre au système. Dans ce cas, qui rappelle les liens entre l'intelligence artificielle et l'interaction homme-machine, l'évaluation seule est plus pertinente. Pourtant, pour se conformer aux évaluations existantes dans le domaine, les résultats présentés dans cette partie ont été obtenus avec l'évaluation complète.

8 Conclusion

Cet article présente une approche nouvelle pour la détection en ligne de nouveauté, qui vise à découvrir automatiquement des concepts - les événements - rattachés à un flux de documents textuels. Cette tâche doit être capable d'apprendre de manière non-supervisée des modèles d'événements qui d'une part évoluent dans le temps et d'autre part sont bruités, privilégiant des solutions robustes. Pour cela, nous avons proposé un formalisme probabiliste simple de la tâche et visant à mieux intégrer l'information temporelle dont on dispose sur les documents. Notre formalisme permet également d'intégrer la notion de *nouveauté* au coeur même du modèle. Nos résultats expérimentaux démontrent le bien fondé de cette approche. Enfin, l'intégration d'une sélection de caractéristiques permet de diminuer la complexité, tout en améliorant la robustesse du système. Pourtant, comme l'ont rappelé les dernières publications du domaine, la tâche demeure un challenge ouvert dans l'attente d'un saut significatif de performances. Les principales perspectives concernent l'intégration dans notre formalisme probabiliste des approches hiérarchiques [Yang et al., 2002] et de l'intégration de modèles probabilistes de séquence [Binsztok, 2003] permettant la découverte de motifs fortement spécifiques à un événement comme par exemple "tremblement [de] terre [de] Boumerdès", "immunité [du] président".

Références

[Allan et al., 1998] Allan J., Carbonell J., Doddington G., Yamron J., and Yang Y. (1998). Topic detection and tracking pilot study final report.

[Allan et al., 1999] Allan J., Jin H., Rajman M., Wayne C., Gildea D., Lavrenko V., Hoberman R., and Caputo D. (1999). Topic-based novelty detection : 1999 summer workshop at clsp.

[Baker et al., 1999] Baker D., Hofmann T., McCallum A., and Yang Y. (1999). A hierarchical probabilistic model for novelty detection in text. In *ICML-99*.

[Binsztok, 2003] Binsztok H. (2003). Suivi d'information et modélisation hiérarchique de séquences. Rapport de Présoutenance LIP6.

[Binsztok and Gallinari, 2002] Binsztok H. and Gallinari P. (2002). Un algorithme en-ligne pour la détection de nouveauté dans un flux de documents. In *JADT - ICTDSA*, volume 6, Saint-Malo, France.

[Cover and Thomas, 1991] Cover T. M. and Thomas J. A. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications.

[Cutting et al., 1992] Cutting D., Karger D., Pedersen J., and Tukey J. (1992). Gather : A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, pages 318–329, Copenhagen, Denmark.

[Kittler, 1974] Kittler J. (1974). Mathematical methods of feature selection in pattern recognition. Technical report, Cambridge University Engineering Dept.

[Manning and Schütze, 1999] Manning C. D. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

[McCallum and Nigam, 1998] McCallum A. and Nigam K. (1998). A comparison of event models for naive bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48.

[Parra et al., 1996] Parra L., Deco G., and Miesbach S. (1996). Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation*, 8(2) :260–269.

[Rissanen, 1982] Rissanen J. (1982). A universal prior for integers and estimation by Minimum Description Length. *Annals of Statistics*, 11 :416–431.

[Robertson et al., 1995] Robertson S., Walker S., and Jones S. (1995). Okapi at trec-3. In 3rd Annual Text REtrieval Conference. NIST.

[Salton, 1968] Salton G. (1968). *Automatic information organization and retrieval*. McGraw-Hill, New York.

[Spitters and Kraaij, 2001] Spitters M. and Kraaij W. (2001). Tno at tdt2001 : Language model-based topic detection. In *TDT*.

[Stokes and Carthy, 2001] Stokes N. and Carthy J. (2001). First story detection using a composite document representation. In *In the proceedings of HLT*, San Diego, CA.

[Walls et al., 1999] Walls F., Jin H., Sista S., and Schwartz R. (1999). Topic detection in broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*, pages 193–198.

[Wayne, 1998] Wayne C. (1998). Topic detection and tracking (tdt) overview and perspective. In *DARPA Broadcast News Transcription and Understanding Workshop*.

[Yang et al., 2000] Yang Y., Ault T., Pierce T., and Lattimer C. W. (2000). Improving text categorization methods for event tracking. In Belkin N. J., Ingwersen P., and Leong M.-K. editors, *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 65–72, Athens, GR. ACM Press, New York, US.

[Yang et al., 1999] Yang Y., Carbonell J., Brown R., Pierce T., Archibald B., and Liu X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4).

[Yang et al., 1998] Yang Y., Pierce T., and Carbonell J. (1998). A study on retrospective and on-line event detection. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 28–36, Melbourne, AU.

[Yang et al., 2002] Yang Y., Zhang J., Carbonell J., and Jin C. (2002). Topic-conditioned novelty detection. In *ACM SIGKDD Internaltional Conference on Knowledge Discovery and Data Mining*, pages 688–693.