Étiquetage en rôles événementiels fondé sur l'utilisation d'un modèle neuronal

Emanuela Boroş[1,2] Romaric Besançon[1] Olivier Ferret[1] Brigitte Grau[2, 3]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, F-91191, Gif-sur-Yvette
(2) LIMSI, Rue John von Neumann, Campus Universitaire d'Orsay, F-91405 Orsay cedex
(3) ENSIIE, 1 square de la résistance F-91025 Évry cedex
{emanuela.boros,romaric.besancon,olivier.ferret}@cea.fr, brigitte.grau@limsi.fr

Résumé. Les systèmes d'extraction d'information doivent faire face depuis toujours à une double difficulté : d'une part, ils souffrent d'une dépendance forte vis-à-vis du domaine pour lesquels ils ont été développés ; d'autre part, leur coût de développement pour un domaine donné est important. Le travail que nous présentons dans cet article se focalise sur la seconde problématique en proposant néanmoins une solution en relation avec la première. Plus précisément, il aborde la tâche d'étiquetage en rôles événementiels dans le cadre du remplissage de formulaire (*template filling*) en proposant pour ce faire de s'appuyer sur un modèle de représentation distribuée de type neuronal. Ce modèle est appris à partir d'un corpus représentatif du domaine considéré sans nécessiter en amont l'utilisation de prétraitements linguistiques élaborés. Il fournit un espace de représentation permettant à un classifieur supervisé traditionnel de se dispenser de l'utilisation de traits complexes et variés (traits morphosyntaxiques, syntaxiques ou sémantiques). Par une série d'expérimentations menées sur le corpus de la campagne d'évaluation MUC-4, nous montrons en particulier que cette approche permet de dépasser les performances de l'état de l'art et que cette différence est d'autant plus importante que la taille du corpus d'entraînement est faible. Nous montrons également l'intérêt de l'adaptation de ce type de modèle au domaine traité par rapport à l'utilisation de représentations distribuées à usage générique.

Abstract. Information Extraction systems must cope with two problems : they heavily depend on the considered domain but the cost of development for a domain-specific system is important. We propose a new solution for role labeling in the event-extraction task that relies on using unsupervised word representations (*word embeddings*) as word features. We automatically learn domain-relevant distributed representations from a domain-specific unlabeled corpus without complex linguistic processing and use these features in a supervised classifier. Our experimental results on the MUC-4 corpus show that this system outperforms state-of-the-art systems on this event extraction task, especially when the amount of annotated data is small.We also show that using word representations induced on a domain-relevant dataset achieves better results than using more general word embeddings.

Mots-clés : Extraction d'information, extraction de rôles événementiels, modèles de langage neuronaux.

Keywords: Information extraction, event role filler detection, neural language models.

1 Introduction

Un enjeu majeur de l'Extraction d'Information (EI) consiste à aider un utilisateur à identifier rapidement des événements ainsi que leurs entités descriptives dans de très grands volumes de documents. L'extraction d'événements peut porter sur des domaines variés. Dans le domaine médical et biologique, la notion d'événement est utilisée pour désigner par exemple le changement d'état d'une molécule en biologie, ou encore l'ensemble des informations concernant l'administration d'un traitement en médecine (Cohen *et al.*, 2009; Yakushiji *et al.*, 2001; Chun *et al.*, 2005). Dans le domaine de l'économie et la finance, les centres d'intérêt concernent les fusions, acquisitions et échanges d'entreprises ou de produits (Hung *et al.*, 2010; Michaely *et al.*, 1995).

Un événement est décrit par un ensemble de participants (*i.e.* des attributs ou rôles) dont les valeurs sont des extraits de texte, correspondant à des entités nommées ou des entités du domaine. Par exemple, un acte terroriste est un événement dont les participants sont les auteurs, les victimes ou encore les cibles. En domaine biomédical, un type d'événement largement étudié est celui des interactions où les rôles désignent des protéines ou des gènes, des médicaments ou autres

molécules. Cette problématique est issue des campagnes d'évaluation MUC [1] (Grishman & Sundheim, 1996), TREC (Voorhees & Tong, 2011) et ACE (Strassel *et al.*, 2008) qui ont fortement contribué à l'évolution du domaine. Dans cet article, nous nous intéressons plus spécifiquement à la reconnaissance des entités et leur étiquetage en rôle. Cette tâche est complexe et recouvre des problématiques telles que la reconnaissance d'entités nommées, la reconnaissance de rôles sémantiques ou l'extraction de relations binaires.

Pour cette tâche, beaucoup de systèmes proposent des méthodes d'extraction de patrons ou de génération de règles fondées sur le contexte environnant, local et global (Patwardhan & Riloff, 2009; Huang & Riloff, 2011). Les méthodes d'acquisition de ces patrons incluent des approches par amorçage (Huang & Riloff, 2012a; Yangarber *et al.*, 2000), de l'apprentissage faiblement supervisé (Huang & Riloff, 2011; Sudo *et al.*, 2003; Surdeanu *et al.*, 2006), de l'apprentissage supervisé (Chieu *et al.*, 2003; Freitag, 1998; Bunescu & Mooney, 2004; Patwardhan & Riloff, 2009), et autres variations. Les patrons appris par ces approches sont ensuite utilisés pour reconnaître et étiqueter, dans de nouveaux documents, des extraits de textes en tant que valeurs d'attributs.

Toutes ces méthodes reposent sur une part assez importante d'annotations manuelles couplée à l'utilisation intensive de connaissances linguistiques et les performances obtenues sont donc en rapport avec la possibilité de mettre en œuvre cette masse de connaissances ainsi que la capacité à définir les ensembles de traits en entrée des classifieurs. De plus, la bonne application de ces méthodes nécessite de connaître a priori le domaine d'application. Il devient ainsi difficile d'appliquer efficacement une méthode donnée sur un domaine différent.

Dans ce travail, nous abordons la tâche d'étiquetage d'entités en rôles décrivant un événement, que nous nommons étiquetage en rôles événementiels, par l'apprentissage automatique de traits pertinents qui ne nécessite qu'un nombre limité de connaissances préalables. A cette fin, des représentations de mots (*word embeddings*) sont induites par application non supervisée d'un réseau de neurone comme dans (Bengio *et al.*, 2006; Collobert & Weston, 2008) sur des données brutes. Les valeurs d'attributs relatives aux exemples connus d'événements à extraire sont annotées dans des textes et transformées à partir des représentations apprises pour entraîner un classifieur permettant de prédire l'étiquette du rôle rempli. Notre objectif est double : (1) montrer que des représentations de mots apprises de façon non supervisée ont une capacité de généralisation et de représentation du sens qui les rend compétitives sur la tâche d'étiquetage en rôles événementiels, (2) montrer que ces représentations sont évolutives et robustes lorsqu'on fait varier la taille des données d'apprentissage.

Nous avons évalué notre approche sur les données issues de MUC-4 (Lehnert *et al.*, 1992) qui portent sur des actes terroristes, et donc sur la reconnaissances des auteurs, victimes et cibles, et nous obtenons des résultats supérieurs à ceux des méthodes état de l'art sur ces mêmes données (Huang & Riloff, 2011, 2012a; Patwardhan & Riloff, 2009).

Après avoir présenté l'état de l'art en extraction d'événements et en apprentissage de représentations dans la section 2, la suite de cet article décrit plus précisément notre approche en section 3. Les données d'expérimentation et les résultats de l'évaluation sont ensuite présentés dans la section 4.

2 État de l'art

Bien qu'il n'existe pas une manière unique d'aborder l'extraction d'événements à partir de textes, celle-ci est reconnue comme un problème complexe que l'on décompose en différentes tâches prototypiques [2] : détection des mentions d'événement ; extraction des candidats au remplissage des rôles ; rattachement local, souvent au niveau phrastique, des candidats aux mentions d'événement ; fusion au niveau textuel des candidats au remplissage des rôles. Le problème que nous abordons ici est celui de la détection des candidats au remplissage des rôles d'un événement, tâche que l'on peut considérer également comme une annotation de phrases en rôles événementiels. Les candidats sont en toute généralité des groupes nominaux, dont certains peuvent correspondre à des entités nommées.

Deux grands types d'approches ont été proposés pour la tâche d'extraction d'événements : les approches fondées sur l'application de patrons (Krupka *et al.*, 1991; Hobbs *et al.*, 1992; Riloff, 1996a,b; Yangarber *et al.*, 2000)) et les approches par apprentissage (Chieu *et al.*, 2003; Freitag, 1998; Huang & Riloff, 2011; Patwardhan & Riloff, 2009; Yangarber *et al.*, 2000; Surdeanu *et al.*, 2006)).

Les patrons sont acquis à partir de textes par application de règles reposant sur des connaissances syntaxiques, extraites d'un arbre syntaxique par exemple, et sémantiques pour identifier les rôles. Les premiers systèmes (Krupka *et al.*, 1991;

1. MUC 1-7 Message Understanding Conferences de 1987 à 1998 organisées par le DARPA.
2. La décomposition que nous faisons ici est essentiellement fonctionnelle et ne fait pas apparaître les liens de dépendance pouvant exister entre ces différentes tâches.

Hobbs *et al.*, 1992; Riloff, 1996a) sont issus des conférences MUC. AutoSlog-TS (Riloff, 1996a), utilisé comme système de base dans nos évaluations et qui est une version améliorée de AutoSlog (Riloff, 1996b), propose une séparation en textes pertinents et non pertinents et un ordonnancement des patrons extraits. Le principal inconvénient de ces systèmes est qu'ils font appel à une vérification manuelle pour sélectionner les patrons, qui peut s'avérer coûteuse.

L'intérêt porté aux approches par apprentissage s'est largement développé et a donné lieu à de nouveaux systèmes, qui reposent sur des méthodes d'apprentissage complètement supervisées (Chieu *et al.*, 2003; Freitag, 1998; Bunescu & Mooney, 2004; Patwardhan & Riloff, 2009) ou faiblement supervisées (Huang & Riloff, 2011; Sudo *et al.*, 2003; Surdeanu *et al.*, 2006). La dépendance des systèmes à l'existence d'annotations riches des textes s'est relâchée avec l'apparition des techniques d'amorçage (Huang & Riloff, 2012a; Yangarber *et al.*, 2000).

Par exemple, le système ALICE (Chieu *et al.*, 2003) est fondé sur différents algorithmes d'apprentissage, utilisant un riche ensemble de traits syntaxiques et sémantiques, et montre que ces types de traits permettent d'améliorer les performances des systèmes. Les auteurs de (Bunescu & Mooney, 2004) proposent une variante des CRFs (Conditional Random Fields) pour exploiter les relations entre traits et en montrent l'intérêt sur la tâche d'extraction d'événements biomédicaux.

Différentes approches ont exploré l'importance du contexte environnant pour reconnaître des valeurs de rôles. GLACIER (Patwardhan & Riloff, 2009) va au-delà de l'analyse du contexte local de la mention d'un événement (*i.e.* la proposition) pour analyser un second contexte plus large au niveau de la phrase. En reprenant cette idée, nous explorons aussi un contexte assez large pour rechercher des valeurs de rôles lors de l'apprentissage du modèle de langue. TIER (Huang & Riloff, 2011) a exploré le fait d'écarter un contexte s'il est non pertinent, même si dans certaines situations une phrase non pertinente peut mentionner les suites d'un événement où certains rôles sont précisés. TIER repose sur une suite de traitements héritée des systèmes précédents partant de la reconnaissance de contextes spécifiques à des rôles par différentes couches de classifieurs et finissant, au niveau le plus bas, par l'extraction des valeurs des rôles. L'amélioration de TIER par l'usage de co-références et de relations de discours est étudiée dans (Huang & Riloff, 2012b,a).

De tous ces travaux, on peut voir que les recherches en extraction d'événements vont vers l'ajout de traits riches donnés en entrée de chaînes de classifieurs, et vers une exploration du contexte des entités à étiqueter. Le but de (Huang & Riloff, 2012b) est de pouvoir reconnaître des transitions et des relations de discours dans le texte de manière à pouvoir mieux identifier les contextes d'apparition des rôles pour un événement donné. Les candidats aux rôles liés à un événement sont identifiés indépendamment, par une approche montante, puis les contextes sélectionnés en mettant en oeuvre des connaissances pour déterminer la cohésion textuelle. PIPER (Patwardhan & Riloff, 2007; Patwardhan, 2010) est construit à partir d'une classification des phrases qui distingue les régions pertinentes et non pertinentes et apprend des patrons d'extraction pertinents pour le domaine selon une mesure d'affinité sémantique. On peut aussi ajouter que, même si l'amélioration des performances est réelle, ces ajouts peuvent rendre les systèmes très lents et non utilisables dans des applications à grande échelle. C'est pourquoi TIER$_{light}$ (Huang & Riloff, 2012a) propose de diminuer le recours à des annotations lourdes, pour faciliter le passage d'un domaine à un autre, par l'application de techniques d'amorçage pour l'étiquetage en rôles.

Notre approche partage avec ces approches l'importance donnée au contexte des mots impliqués dans des valeurs de rôles. Cependant, alors que ces systèmes reposent sur la conception de riches ensembles de traits à donner en entrée des classifieurs, notre approche réduit cette complexité en donnant seulement les mots bruts en entrée d'un réseau de neurones. Les traits sont appris automatiquement et sont réutilisés dans la tâche d'étiquetage en rôles événementiels ; ils permettent de plus d'obtenir de meilleurs résultats à partir de ces seules données. Ce type d'approche, proposé dans des travaux comme (Bengio *et al.*, 2006; Collobert & Weston, 2008; Turian *et al.*, 2010), a montré des résultats intéressants sur de nombreuses tâches en traitement automatique des langues mais n'a jamais été appliqué à l'extraction d'événements.

3 Méthode

3.1 Principes

À l'instar de (Huang & Riloff, 2012b), la tâche d'étiquetage en rôles événementiels est réalisée comme une tâche indépendante des autres tâches mentionnées section 2. Son objectif est de produire un ensemble assez large de candidats qui seront ensuite filtrés par les contraintes de rattachement aux événements, soit au niveau local, soit au niveau global. À la différence de (Jean-Louis *et al.*, 2011), nous ne faisons pas l'hypothèse que ces candidats se limitent à des entités nommées et nous nous ne faisons pas non plus l'hypothèse d'une bijection entre le rôle d'un événement et un type d'entité nommée.

Sur le plan méthodologique, nous traitons cette tâche d'étiquetage en rôles événementiels sous l'angle de la classification supervisée. Nous nous appuyons pour ce faire sur la sortie d'un outil générique de découpage des phrases en chunks syntaxiques et nous appliquons un classifieur multiclasse à chaque chunk nominal extrait pour déterminer à quel rôle du type d'événement considéré il est susceptible de se rattacher. Nous avons ainsi une classe par rôle à laquelle s'ajoute une classe correspondant à l'absence de rattachement. L'originalité de l'approche que nous proposons réside dans le type de représentation des chunks nominaux exploité par notre classifieur. Pour ce type de tâche, il est habituel de représenter chaque candidat à un rôle par un ensemble de traits caractérisant différents types d'informations allant des simples mots le constituant jusqu'à son rôle sémantique dans la phrase en passant par la catégorie morphosyntaxique de ses constituants ou son rôle syntaxique. Comme nous l'avons mentionné dans la section précédente, cette approche a un triple inconvénient : elle nécessite un ensemble d'outils élaborés qui ne sont pas toujours disponibles pour une langue donnée ; ces outils n'étant pas parfaits, les informations qu'ils délivrent sont entachées d'un certain taux d'erreur, qui a tendance à être d'autant plus conséquent que l'outil est plus élaboré ; enfin, ces outils sont génériques et donc, la plupart du temps, non adaptés au domaine considéré.

Pour faire face à ces problèmes, nous proposons d'adopter une approche différente, inspirée de travaux tels que (Collobert & Weston, 2008), consistant à projeter les candidats à un rôle événementiel, à partir de leurs mots, dans un espace de représentation défini spécifiquement pour le domaine considéré. Plus précisément, cet espace est construit grâce à un réseau de neurones en reprenant des techniques développées pour l'apprentissage de modèles de langage (Bengio *et al.*, 2003). Outre leur adaptation au domaine, les représentations ainsi élaborées ont l'avantage de pouvoir être comparées et leur proximité dans cet espace est à mettre en relation avec la proximité de leur rôle vis-à-vis du domaine. Une fois construites, ces représentations sont utilisées comme traits dans un classifieur supervisé réalisant l'étiquetage en rôles événementiels, à l'instar des traits habituellement utilisés pour cette tâche.

Nous commençons par détailler la façon dont ces représentations sont construites sur un plan générique avant de préciser leur utilisation et les stratégies mises en œuvre pour les adapter à notre contexte de travail.

3.2 Construction des représentations lexicales distribuées

Notre construction de représentations lexicales distribuées (*word embeddings*) s'appuie sur les principes définis dans (Collobert & Weston, 2008). Ces principes sont eux-mêmes issus de la problématique des modèles de langage neuronaux (Bengio *et al.*, 2003). Dans ce contexte, un réseau de neurones est entraîné à prédire la probabilité d'un mot en fonction de son contexte. L'entraînement d'un tel modèle passe par le traitement d'un large ensemble d'exemples de séquences de mots et une optimisation des paramètres du réseau du point de vue de sa capacité à fournir les meilleures prédictions pour les séquences exemples.

Une des spécificités des réseaux utilisés réside dans la représentation des séquences de mots qu'ils prennent comme entrée et plus spécifiquement des mots composant ces séquences. Dans ce schéma de représentation en effet, un mot n'est plus considéré comme un simple symbole mais possède une représentation distribuée. Cette représentation prend la forme d'un ensemble fixe de dimensions valuées, c'est-à-dire un vecteur de nombres réels de même taille pour tous les mots du vocabulaire considéré. De ce point de vue, cette représentation est proche de représentations issues de méthodes de réduction de dimensions telles que celles produites par l'Analyse Sémantique Latente par exemple (Landauer *et al.*, 1998). À la différence de ces méthodes de réduction de dimensions, qui appliquent une transformation mathématique donnée, les représentations produites dans le cadre des modèles de langage neuronaux sont apprises en relation avec les exemples exploités. Elles sont donc intrinsèquement adaptées à ces derniers. (Collobert & Weston, 2008) a repris ce schéma mais avec la perspective plus générale de construire des représentations lexicales distribuées utilisables pour des tâches autres que la prédiction de la probabilité d'une séquence de mots.

Nous nous inscrivons dans le prolongement de (Collobert & Weston, 2008) pour construire des représentations dédiées à l'étiquetage en rôles événementiels. Cette construction prend la forme de l'apprentissage d'un modèle permettant de différencier de façon générique une séquence de mots issue d'un corpus représentatif d'un domaine et une séquence de mots proche mais ne figurant pas dans le corpus. En pratique, les secondes sont construites par l'altération des premières en changeant un de leurs mots, en l'occurrence celui du milieu. Nous verrons à la section suivante comment s'effectue ce changement. La tâche peut donc être vue comme un test de compatibilité du mot central d'une séquence avec son contexte environnant du point de vue du corpus considéré et donc, de son domaine associé.

Le modèle à apprendre prend plus spécifiquement la forme d'un réseau de neurones à trois couches, comme l'illustre la figure 1, avec une première couche (de gauche à droite) permettant de représenter les séquences en entrée et une couche

finale ayant pour rôle de leur attribuer un score. Les séquences correspondent au contenu d'une fenêtre glissante déplacée sur les textes et composée de $m = 2n+1$ mots. Chaque unité de la couche d'entrée du réseau de la figure 1 ne correspond pas directement à un mot de cette fenêtre mais à une des k dimensions de sa représentation distribuée. La couche d'entrée du réseau est ainsi formée de la concaténation des représentations distribuées des m mots de la fenêtre et contient donc $k \cdot m$ unités.

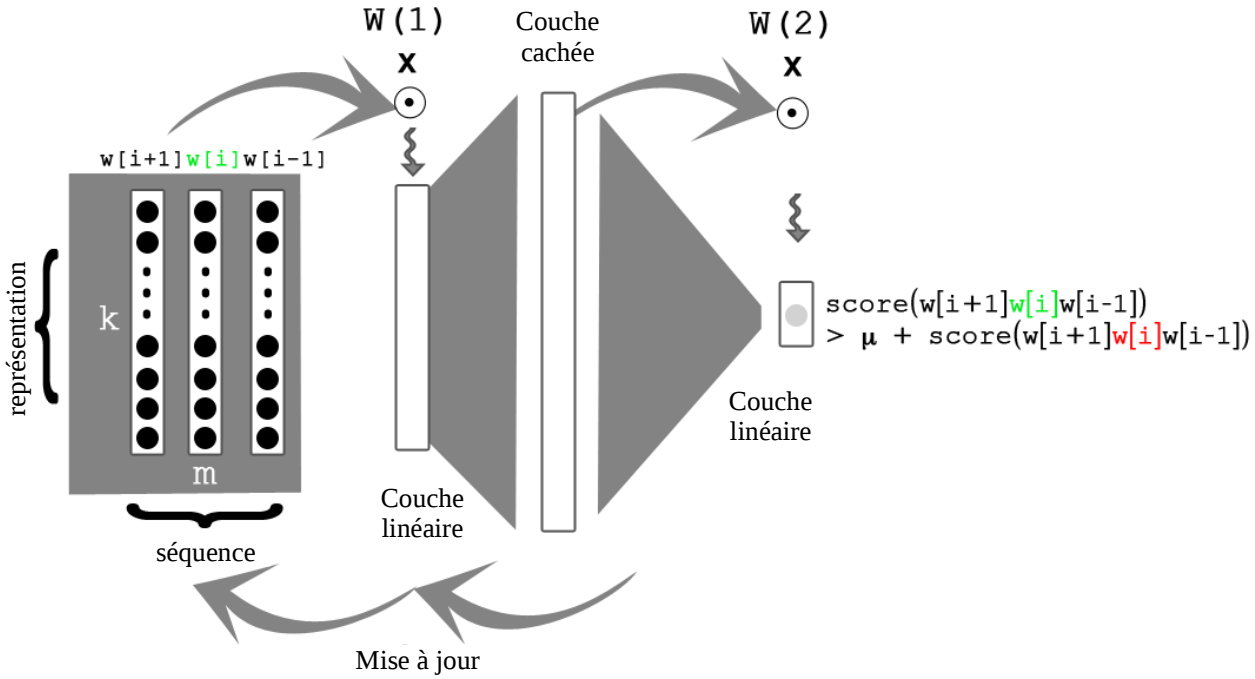

FIGURE 1 – Architecture du réseau de neurones utilisé

Lors de la première phase du processus d'apprentissage, les différentes dimensions de la représentation d'un mot sont initialisées de manière aléatoire selon une loi uniforme. L'activation correspondante est propagée dans le réseau, d'abord vers la couche cachée, puis vers la couche de sortie pour aboutir au calcul d'un score associé à la séquence d'entrée. D'un point de vue plus formel, pour la séquence d'entrée $\langle w_i \rangle = \langle w_{i-n)} \dots, w_{i-1}, w_i, w_{i+1} \dots w_{i+n} \rangle$, on a ainsi :

$$\begin{aligned} score(\langle w_i \rangle) &= b^{(2)} + W^{(2)} h(\langle w_i \rangle) \\ h(\langle w_i \rangle) &= \Phi(b^{(1)} + W^{(1)} \langle w_i \rangle) \end{aligned} \tag{1}$$

$b^{(1)}$, $b^{(2)}$ étant les termes de biais, sous la forme de vecteurs, intervenant dans le cumul des activations en entrée d'une unité, $W^{(1)}$ et $W^{(2)}$, les matrices de poids des connexions entre couches et Φ, la fonction d'activation de la couche cachée. Dans cette configuration, cette fonction est non linéaire, avec le choix dans notre de cas de $softsign(a) = |a|/(1+|a|)$ qui présente l'avantage de permettre des temps d'apprentissage réduits.

La différence entre le score ainsi calculé pour une séquence véritablement observée et une séquence altérée en changeant son mot central est utilisée comme critère d'optimisation pour la mise à jour à la fois des poids des connexions du réseau et de la valeur des dimensions de la représentation des mots de la séquence d'entrée. Cette mise à jour est réalisée de façon classique par l'application d'une descente de gradient stochastique. Le critère d'optimisation est donc un critère d'ordonnancement, à la différence des travaux antérieurs à (Collobert & Weston, 2008), qui optimisaient la log-vraisemblance pour le dernier mot de la séquence et devaient donc évaluer en sortie une probabilité pour tous les mots du vocabulaire pour chaque séquence en entrée. Plus formellement, ce critère d'ordonnancement spécifie que le score d'une séquence observée $\langle w_i \rangle$ doit être plus grand que celui de tout autre séquence $\langle \tilde{w_j} \rangle$ produite par remplacement du mot central de la séquence

observée par un autre mot du dictionnaire, et ceci avec une marge de μ, donc tel que $score(\langle w_i \rangle) > \mu + score(\langle \tilde{w_i} \rangle)$, où $\mu = 0,1$ comme dans (Collobert & Weston, 2008).

Au final, la représentation modifiée par ce critère de chaque mot de la séquence d'entrée est ensuite stockée pour être réutilisée lorsqu'une autre séquence contenant ce mot est présentée en entrée du réseau. Les représentations des mots sont ainsi adaptées de façon incrémentale en fonction du critère d'optimisation retenu.

3.3 Stratégie d'adaptation des représentations distribuées à la tâche

À la section précédente, nous avons défini la façon dont sont construites les représentations lexicales distribuées pour l'étiquetage en rôles événementiels, en reprenant fortement les principes définis dans (Collobert & Weston, 2008). Nous avons néanmoins réalisé une modification spécifique de cette méthode pour une meilleure adaptation des représentations construites à notre tâche. L'idée sous-jacente à cette modification est de favoriser, dans les exemples fournis pour l'entraînement du modèle, la présence de mots importants du domaine, de telle sorte que l'apprentissage s'effectue plus rapidement.

Pour évaluer l'importance d'un mot par rapport à un domaine, lequel s'identifie dans notre cas à un type d'événement, nous adoptons une approche faiblement supervisée en évaluant la proximité sémantique entre ce mot et un ensemble de mots représentatifs des événements considérés, appelés *étiquettes événements*. Par exemple, dans le cas du corpus MUC que nous avons utilisé pour nos expérimentations de la section 4, les événements sont des attaques terroristes et les mots choisis pour les représenter sont les étiquettes événements $\{attack, bombing, kidnapping, arson\}$. La proximité sémantique entre un mot et une étiquette événement est définie par la mesure de *Leacock Chodorow*. Cette mesure de similarité lexicale se fonde sur WordNet, en l'occurrence sa version 3.0, et dépend de la longueur du chemin le plus court entre deux synsets dans la hiérarchie de WordNet, normalisée par la hauteur de cette hiérarchie. Plus formellement, elle s'écrit : $-log(p/2 \cdot D)$ où p est la longueur du chemin entre les synsets des mots considérés et D est la hauteur de la hiérarchie de WordNet. La proximité d'un mot par rapport à un domaine est ainsi donnée par la valeur moyenne de la mesure de *Leacock Chodorow* entre ce mot et chacun des mots événements du domaine.

Pour favoriser la présence des mots importants du domaine dans les exemples, nous choisissons de modifier la stratégie de sélection du mot remplaçant le mot central d'une séquence exemple lors de la corruption de cette séquence. Dans (Collobert & Weston, 2008), ce choix est aléatoire parmi la totalité des mots du vocabulaire pris en compte pour construire les représentations. Dans notre cas, nous utilisons la méthode d'évaluation de l'importance d'un mot par rapport au domaine présentée ci-dessus pour ordonner les mots du vocabulaire et choisir le mot remplaçant de façon aléatoire parmi les mots ayant un score supérieur à un seuil donné.

3.4 Utilisation des représentations distribuées pour l'étiquetage en rôles événementiels

Les représentations apprises pour chacun des mots permettent de calculer les traits des exemples donnés en entrée du classifieur supervisé en vue de prédire leur étiquette, *i.e.* leur rôle événementiel. Néanmoins, les rôles événementiels ne sont en général pas occupés par de simples mots mais plutôt par des groupes nominaux, pouvant s'identifier dans certains cas à des entités nommées. Pour l'identification de ces rôles événementiels, nous avons donc opéré en deux temps. En premier lieu, nous avons appliqué un analyseur en chunks pour identifier les candidats à ces rôles. En l'occurrence, tout chunk nominal est considéré comme un candidat, les constituants des autres chunks recevant une étiquette NULL (cf. section 4.2).

Dans un second temps, nous avons appliqué un classifieur préalablement entraîné sur un corpus annoté pour décider quel rôle, s'il en occupe un, un chunk occupe pour le type d'événement considéré. Pour ce faire, il est nécessaire de passer de la représentation construite pour chaque mot à la représentation d'un chunk. Ce passage est réalisé via le mécanisme du *max-pooling*. Un chunk de N mots est ainsi représenté avec le même nombre de dimensions qu'un mot et chacune de ses dimensions i prend pour valeur $max(w_{i1}, ..., w_{iN})$ où w_{ij} est la valeur de la dimension i pour le mot w_j constituant le chunk.

Pour la classification proprement dite, nous nous appuyons sur la variante *Extra-Trees* (Geurts *et al.*, 2006) des forêts d'arbres décisionnels telle qu'elle est implémentée dans (Pedregosa *et al.*, 2011).

4 Expérimentations et résultats

4.1 Description de la tâche

Nous avons évalué le système présenté sur les données de la campagne d'évaluation MUC-4, qui forment un corpus d'évaluation standard pour la tâche d'extraction d'événement. Le corpus d'entraînement comporte 1 500 textes et modèles d'événements (*template*) instanciés associés. La tâche consiste à extraire les informations descriptives d'événements terroristes en Amérique Latine. Étant donné un texte, il s'agit de remplir une structure pour chaque événement décrit (par exemple attaque, enlèvement, prise d'otage, pose de bombe, etc.). Si le texte décrit plus d'un événement, il faut remplir une structure pour chacun d'eux. Les tests officiels, nommés TST3 et TST4, contiennent 100 documents chacun provenant de cet ensemble. Nous avons entraîné notre système sur 1300 documents et l'avons testé à chaque fois sur le même ensemble de test, formé de la conjonction des deux ensembles de tests TST3+TST4.

Une modèle d'événement comporte un ensemble d'attributs prédéfinis correspondant aux valeurs qui doivent être trouvées dans les textes, (dans les modèles d'événements de MUC-4, il y a 25 attributs). Ces attributs sont de types différents qui nécessitent d'être traités différemment, les valeurs de ces attributs devant être extraites ou inférées à partir des textes [3]. Ces attributs peuvent être divisés en trois catégories :

1. les attributs de type texte : ces attributs sont remplis par des chaînes de caractères extraites directement des textes (6. INCIDENT : INSTRUMENT ID, 9. PERP : INDIVIDUAL ID, 10. PERP : ORGANIZATION ID, 12. PHYS TGT : ID, 18. HUM TGT : NAME, 19. HUM TGT : DESCRIPTION, 6. INCIDENT : INSTRUMENT ID, 7. INCIDENT : INSTRUMENT TYPE). Ils ne correspondent pas forcément à une entité nommée ;
2. les attributs calculés : les valeurs doivent être calculées à partir d'extraits de textes. Par exemple, INCIDENT : DATE doit être inférée d'expressions temporelles telles que *today, last week,* etc.
3. les attributs à valeur contrainte : la valeur de ce type d'attribut provient d'un ensemble fini de valeurs possibles. Elles doivent souvent être inférées des documents.

Pour notre évaluation, nous nous sommes concentrés sur l'instanciation des attributs texte, de façon similaire aux autres systèmes de l'état de l'art. De façon similaire à (Patwardhan & Riloff, 2009), nous distinguons cinq grands groupes d'attributs :

AutInd	(PERP :INDIVIDUAL ID)	*AutOrg*	(PERP :ORGANIZATION ID)
Cible	(PHYS TGT :ID)	*Victime*	(HUM TGT :NAME, HUM TGT :DESCRIPTION)
Arme	(INCIDENT :INSTRUMENT ID, INCIDENT :INSTRUMENT TYPE)		

Nous évaluons ensuite la précision des extractions. Notons que, comme dans les travaux comparables (Patwardhan & Riloff, 2009; Huang & Riloff, 2010, 2011, 2012b), nous ne nous intéressons pas à la construction complète des structures événementielles mais seulement à l'identification des rôles événementiels, quel que soit l'événement auquel ils sont reliés. Pour établir la correspondance entre les valeurs extraites et les valeurs de référence, on compare les têtes des chunks (l'extraction de *men* est considérée correcte pour une réponse attendue de *five armed men*), et on fusionne les extractions multiples (de sorte que plusieurs chunks extraits partageant la même tête ne sont comptés qu'une seule fois). Enfin, cette évaluation prend en compte les rôles multi-valués, en distinguant les conjonctions (lorsque plusieurs victimes sont nommées, on doit les trouver toutes) et les disjonctions (lorsque la même entité a plusieurs noms, il suffit d'en trouver un seul).

4.2 Étiquetage du corpus pour l'apprentissage supervisé

Comme indiqué dans la section 3.4, nous utilisons un classifieur permettant d'associer chaque chunk du texte à un rôle événementiel. Ce classifieur doit donc être entraîné sur un corpus annoté correspondant à cette tâche, qui est construit automatiquement à partir des événements de référence. Les valeurs des attributs sont retrouvées dans les documents correspondants, en appliquant un seuil de distance minimale pour aligner les mots. Par exemple, si un attribut a trois valeurs possibles, chacune étant formée de plusieurs mots, nous recherchons tous les mots dans le texte. Si les positions des différents mots composant une valeur sont suffisamment proches, on attribue l'étiquette du rôle à l'empan délimité par ces mots.

3. A l'exception d'attributs de méta-données comme les attributs 0 (MESSAGE :ID) et 1 (MESSAGE : TEMPLATE).

Cet étiquetage automatique a été réalisé en fonction des syntagmes (*i.e.* les chunks) proposés par l'outil SENNA ("Semantic/syntactic Extraction using a Neural Network Architecture", (Collobert *et al.*, 2011)), où à chaque mot est attribué un tag unique, soit mot simple (S-NP), début de chunk (B-NP), interne à un chunk (I-NP) ou fin de chunk (E-NP). Nous associons à l'attribut le plus petit chunk englobant sa valeur. Toutes les variantes des modèles instanciés sont prises en compte. Les groupes restants, qui ne couvrent aucune valeur d'attribut, sont associés à une étiquette NULL. Un exemple de phrase annotée de cette façon est fourni ci-dessous.

phrase initiale	Guerrillas	attacked	the	Santo	Tomas	presidential	farm
chunks (SENNA)	S-NP	S-VP	B-NP	I-NP	I-NP	I-NP	E-NP
rôles événementiels	S-AGENT	NULL	B-TARGET	I-TARGET	I-TARGET	I-TARGET	E-TARGET

4.3 Expérimentations

Après l'annotation automatique du corpus et une normalisation de base (passage du corpus en minuscules, suppression des espaces en trop, découpage en phrases), les représentations lexicales sont apprises en appliquant le réseau de neurones présenté à la figure 1.

Après expérimentations, nous avons retenu des représentations lexicales formées par des vecteurs à 50 dimensions obtenus par application sur des séquences de 5 mots, dénommés *DRVR-50* (pour *Domain-Relevant Vector Representations*. Comme indiqué à la section 3.2, nous avons utilisé le réseau de neurones avec *softsign* comme fonction non linéaire. Vu la faible complexité de cette fonction, la durée d'entraînement est rapide (environ 12 heures), en comparaison des semaines mentionnées dans (Turian *et al.*, 2010).

Nous avons effectué un certain nombre d'expérimentations, que nous ne détaillerons pas ici, afin de déterminer la meilleure combinaison des paramètres importants de notre système. Parmi ceux-ci, nous avons accordé une attention toute particulière à la méthode de corruption des séquences en considérant trois conditions de choix aléatoire du mot corrupteur : choix parmi tout le vocabulaire, choix parmi les mots les plus fréquents et choix parmi les mots les plus liés au domaine selon le critère présenté à la section 3.3. Les meilleurs résultats ont été obtenus avec cette dernière condition, montrant ainsi l'intérêt du mécanisme d'adaptation faiblement supervisé au domaine que nous avons proposé pour la construction

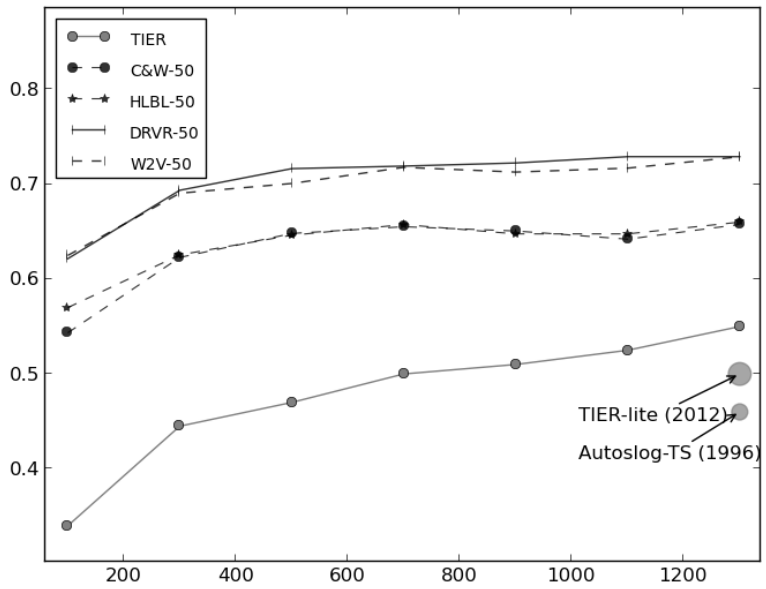

FIGURE 2 – F1-Mesure pour les rôles texte de TST3+TST4 avec différents paramètres, en relation avec la courbe d'apprentissage de TIER (Huang & Riloff, 2012a). Les points gris représentent des résultats marquant de la tâche.

TST3 + TST4						
Approches faiblement supervisées						
	AutInd	**AutOrg**	**Cible**	**Victime**	**Arme**	**Moyenne**
Autoslog-TS (1996)	33/49/40	53/33/41	54/59/56	49/54/51	38/44/41	45/48/46
Piper$_{Best}$ (2007)	39/48/43	55/31/40	37/60/46	44/46/45	47/47/47	44/36/40
TIER$_{lite}$ (2012)	47/51/47	60/39/47	37/65/47	39/53/45	53/55/54	47/53/**50**
Chambers+Jurafsky (2011)	–	–	–	–	–	44/36/40
Modèles supervisés						
GLACIER (2009)	51/58/54	34/45/38	43/72/53	55/58/56	57/53/55	48/57/52
TIER (2011)	48/57/52	46/53/50	51/73/60	56/60/58	53/64/58	51/62/56
(Huang & Riloff, 2012b)	54/57/56	55/49/51	55/68/61	63/59/61	62/64/63	58/60/**59**
Modèles neuronaux						
C&W-50	80/55/65	64/65/64	76/72/74	53/63/57	85/64/73	68/63/65
HLBL-50	81/53/64	63/67/65	78/72/75	53/63/58	93/64/75	69/62/66
W2V-50	79/57/66	88/71/79	74/72/73	69/75/71	97/65/78	77/68/72
DRVR-50	79/57/66	91/74/81	79/57/66	77/75/76	92/58/81	80/67/**73**

TABLE 1 – Résultats sur les rôles texte de TST3 + TST4 P/R/F1 (Précision/Rappel/F1-Mesure)

des représentations lexicales distribuées.

Pour la tâche d'étiquetage supervisé, nous utilisons comme indiqué précédemment un algorithme de forêts d'arbres décisionnels (Extra-Trees), avec 500 arbres, valeur adoptée pour toutes les expérimentations menées.

Les résultats présentés à la figure 2 sont calculés pour les rôles considérés (AutInd, AutOrg, Cible, Victime, Arme). Nous observons que nos représentations lexicales (*DRVR-50*) surpassent les résultats de l'état de l'art indiqués par les points gris, ce qui montre qu'elles permettent de représenter des informations sémantiques au moins équivalentes pour la tâche sans avoir à ajouter d'autres traits. Par ailleurs, on peut voir qu'elles présentent une bonne stabilité par rapport à la taille du corpus d'apprentissage. La méthode que nous proposons est donc une piste intéressante pour développer rapidement des systèmes d'extraction d'événements sur un nouveau domaine avec peu de données annotées.

La table 1 présente des résultats comparatifs plus détaillés. On peut voir dans ce tableau que nos résultats surpassent ceux des modèles faiblement supervisés (0,73 vs 0,59) et supervisés (0,73 vs 0,56). Les rôles *AutOrg* et *Arme* obtiennent même une très bonne précision, ce qui signifie que pour ces rôles, un filtre supplémentaire pour éliminer les faux candidats n'est pas nécessaire. De façon générale, le compromis entre la justesse de la réponse et le nombre de candidats trouvés devra être étudié plus précisément dans de futurs travaux.

L'aspect innovant de notre système concerne l'utilisation des représentations lexicales apprises par un modèle neuronal. Pour étudier de façon plus poussée l'influence de ces représentations, nous avons comparé nos représentations apprises sur le corpus MUC-4 selon le modèle détaillé en section 3 avec des modèles de représentations lexicales existants : nous avons utilisé les données mises à disposition par (Turian *et al.*, 2010), et plus précisément, selon les modèles de C&W et HLBL [4]. Ces représentations sont construites à partir d'un corpus plus important et plus généraliste d'articles de journaux (corpus Reuters RCV1). Les résultats obtenus avec ces représentations lexicales sont reportés dans la table 1 et montrent que les scores obtenus avec notre modèle restent supérieurs (avec une F1-mesure de 0,72 contre 0,65 et 0,66, due surtout à une meilleure précision). Ceci met en évidence qu'un modèle appris sur un corpus spécifique à un domaine permet d'obtenir de meilleurs résultats, même si ce corpus est de taille beaucoup moins importante (alors qu'il est d'usage de considérer que les modèles neuronaux nécessitent souvent des données d'entraînement importantes).

De façon complémentaire, nous avons comparé notre méthode pour apprendre les représentations lexicales sur le corpus MUC-4 avec la méthode proposée par (Mikolov *et al.*, 2011) [5], en utilisant le même corpus. Les résultats sont présentés dans la table 1 sous le nom *W2V50*. Les résultats obtenus sont alors comparables avec ceux obtenus par notre système (légèrement moins bons), ce qui confirme que l'utilisation d'un corpus spécifique au domaine considéré est bien un atout intéressant.

4. Ces données sont disponibles sur `http://metaoptimize.com/projects/wordreprs`.
5. Son code pour générer les représentations est disponible à l'adresse : `https://code.google.com/p/word2vec`.

5 Conclusions et perspectives

Nous avons présenté une nouvelle approche d'étiquetage en rôles événementiels qui permet de réduire le nombre de traits à concevoir manuellement en utilisant des représentations lexicales distribuées apprises de manière non supervisée. Ces types de représentation sont connus pour être indépendants de la tâche, et nous avons montré que l'on pouvait les utiliser dans une tâche d'extraction d'événements, en obtenant des résultats qui surpassent les résultats actuels sur la même tâche. De plus, les représentations apprises le sont en tenant compte du domaine à analyser et cela contribue à l'amélioration des résultats obtenus. Nous avons aussi montré qu'elles étaient stables sur différentes tailles de corpus d'apprentissage. Un second point important de ces résultats concerne l'adaptation d'un système à un nouveau domaine. Dans notre cas, il suffit de fournir seulement des exemples de valeurs de rôles à étiqueter et un corpus pas nécessairement très important, et on peut développer rapidement un système d'extraction d'information. Aucune définition de nouveaux traits ou étude de leur adaptation au domaine n'est requise. Il reste à vérifier que l'on peut obtenir d'aussi bons résultats sur un domaine différent.

Dans le futur, nous envisageons de tester d'autres architectures de réseau de neurones pour tirer parti d'informations que l'on peut obtenir à partir d'un analyseur tel qu'un analyseur à base de grammaire probabiliste hors-contexte. Nous envisageons aussi d'étendre le système à l'ensemble de la tâche, en considérant toutes les sous-tâches ensembles lors de l'apprentissage de manière à considérer les relations qu'elles entretiennent.

Références

BENGIO Y., DUCHARME R. & VINCENT P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.

BENGIO Y., SCHWENK H., SENÉCAL J.-S., MORIN F. & GAUVAIN J.-L. (2006). Neural probabilistic language models. In D. HOLMES & L. JAIN, Eds., *Innovations in Machine Learning*, volume 194 of *Studies in Fuzziness and Soft Computing*, p. 138–186. Springer Berlin Heidelberg.

BUNESCU R. & MOONEY R. J. (2004). Collective information extraction with relational markov networks. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, p. 438–445.

CHIEU H. L., NG H. T. & LEE Y. K. (2003). Closing the gap : Learning-based information extraction rivaling knowledge-engineering methods. In *41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, p. 216–223.

CHUN H.-W., HWANG Y.-S. & RIM H.-C. (2005). Unsupervised event extraction from biomedical literature using co-occurrence information and basic patterns. In *IJCNLP 2004*, p. 777–786. Springer.

COHEN K. B., VERSPOOR K., JOHNSON H. L., ROEDER C., OGREN P. V., BAUMGARTNER JR W. A., WHITE E., TIPNEY H. & HUNTER L. (2009). High-precision biological event extraction with a concept recognizer. In *Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task*, p. 50–58.

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *25rd International Conference of Machine learning*, p. 160–167 : ACM.

COLLOBERT R., WESTON J., BATTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Approach*, **12**, 2493–2537.

FREITAG D. (1998). Information extraction from HTML : Application of a general machine learning approach. In *AAAI*, p. 517–523.

GEURTS P., ERNST D. & WEHENKEL L. (2006). Extremely randomized trees. *Machine Learning*, **63**(1), 3–42.

GRISHMAN R. & SUNDHEIM B. (1996). Message understanding conference-6 : A brief history. In *COLING 1996*, p. 466–471.

HOBBS J. R., APPELT D., TYSON M., BEAR J. & ISRAEL D. (1992). SRI International : Description of the FASTUS system used for MUC-4. In *4th Conference on Message understanding*, p. 268–275.

HUANG R. & RILOFF E. (2010). Inducing domain-specific semantic class taggers from (almost) nothing. In *48th Annual Meeting of the Association for Computational Linguistics*, p. 275–285.

HUANG R. & RILOFF E. (2011). Peeling back the layers : Detecting event role fillers in secondary contexts. In *ACL 2011*, p. 1137–1147.

HUANG R. & RILOFF E. (2012a). Bootstrapped training of event extraction classifiers. In *13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 286–295.

HUANG R. & RILOFF E. (2012b). Modeling textual cohesion for event extraction. In *26th Conference on Artificial Intelligence (AAAI 2012)*.

HUNG S.-H., LIN C.-H. & HONG J.-S. (2010). Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling. *Expert Systems with Applications*, **37**(1), 341–347.

JEAN-LOUIS L., BESANÇON R. & FERRET O. (2011). Text segmentation and graph-based method for template filling in information extraction. In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, p. 723–731.

KRUPKA G., JACOBS P., RAU L. & IWAŃSKA L. (1991). GE : Description of the NLToolset System as Used for MUC-3. In *3rd Conference on Message understanding*, p. 144–149.

LANDAUER T. K., FOLTZ P. W. & LAHAM D. (1998). An introduction to latent semantic analysis. *Discourse processes*, **25**(2-3), 259–284.

LEHNERT W., CARDIE C., FISHER D., MCCARTHY J., RILOFF E. & SODERLAND S. (1992). University of Massachusetts : MUC-4 test results and analysis. In *4th Conference on Message understanding*, p. 151–158.

MICHAELY R., THALER R. H. & WOMACK K. L. (1995). Price reactions to dividend initiations and omissions : overreaction or drift ? *The Journal of Finance*, **50**(2), 573–608.

MIKOLOV T., KOMBRINK S., BURGET L., CERNOCKY J. & KHUDANPUR S. (2011). Extensions of recurrent neural network language model. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5528–5531.

PATWARDHAN S. (2010). *Widening the field of view of information extraction through sentential event recognition*. PhD thesis, University of Utah.

PATWARDHAN S. & RILOFF E. (2007). Effective information extraction with semantic affinity patterns and relevant regions. In *EMNLP-CoNLL 2007*, p. 717–727.

PATWARDHAN S. & RILOFF E. (2009). A unified model of phrasal and sentential evidence for information extraction. In *2009 Conference on Empirical Methods in Natural Language Processing*, p. 151–160.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

RILOFF E. (1996a). Automatically generating extraction patterns from untagged text. In *AAAI'96*, p. 1044–1049.

RILOFF E. (1996b). An empirical study of automated dictionary construction for information extraction in three domains. *Artificial intelligence*, **85**(1), 101–134.

STRASSEL S., PRZYBOCKI M. A., PETERSON K., SONG Z. & MAEDA K. (2008). Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *LREC 2008*.

SUDO K., SEKINE S. & GRISHMAN R. (2003). An improved extraction pattern representation model for automatic ie pattern acquisition. In *41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, p. 224–231.

SURDEANU M., TURMO J. & AGENO A. (2006). A hybrid approach for the acquisition of information extraction patterns. In *EACL-2006 Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, p. 48–55.

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations : a simple and general method for semi-supervised learning. In *48th international Annual Meeting on Association for Computational Linguistics*, p. 384–394.

VOORHEES E. & TONG R. (2011). Overview of the TREC 2011 medical records track. In *TREC 2011*.

YAKUSHIJI A., TATEISI Y., MIYAO Y. & TSUJII J. (2001). Event extraction from biomedical papers using a full parser. In *Pacific Symposium on Biocomputing*, volume 6, p. 408–419.

YANGARBER R., GRISHMAN R., TAPANAINEN P. & HUTTUNEN S. (2000). Automatic acquisition of domain knowledge for information extraction. In *18th Conference on Computational linguistics (COLING 2000)*, p. 940–946.