HAL
open science

Représentations et modèles en extraction d'événements supervisée

Dorian Kodelja, Romaric Besancon, Olivier Ferret

To cite this version:

Dorian Kodelja, Romaric Besancon, Olivier Ferret. Représentations et modèles en extraction d'événements supervisée. Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017), Jul 2017, Caen, France. hal-01561986

HAL Id: hal-01561986

https://hal.science/hal-01561986v1

Submitted on 2 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Représentations et modèles en extraction d'événements supervisée

D. Kodelja R. Besançon O. Ferret

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Gif-sur-Yvette, F-91191, France.
{dorian.kodelja,romaric.besancon,olivier.ferret}@cea.fr

Résumé

Cet article de synthèse retrace l'historique des techniques d'extraction d'événement à partir de textes. Cette tâche a d'abord été traitée par des règles lexico-syntaxiques puis des classifieurs supervisés utilisant tous deux des représentations complexes, fortement dépendantes du domaine et sujettes à la propagation d'erreurs. Récemment, plusieurs approches neuronales ont amélioré l'état de l'art en mettant en avant la réduction des prétraitements nécessaires et donc des possibilités d'erreur. Finalement, nous nuançons ce postulat en présentant de nouvelles méthodes réintroduisant ces informations linguistiques.

Mots Clef

Extraction d'information, apprentissage supervisé, réseaux de neurones, représentations distribuées.

Abstract

We present in this survey the successive approaches to supervised event extraction from texts. The first rule-based systems and the classical statistical methods use complex and domain-dependent representations that are prone to error propagation. In response to these problems, recent neural network systems using embeddings have linked their success to the absence of the preprocessing steps producing theses errors. We nuance this viewpoint by presenting recent methods reintroducing these linguistics features.

Keywords

Information Extraction, supervised learning, neural networks, distributed representations.

1 Introduction

L'extraction d'information est un champ de recherche dont l'objectif consiste à extraire automatiquement des informations factuelles structurées dans un domaine donné à partir de données textuelles peu ou pas structurées. Les premiers systèmes d'extraction d'informations, développés manuellement pour un besoin précis dans un domaine spécifique, n'étaient cependant absolument pas réutilisables. Que ces documents émanent du domaine biomédical (articles), industriel (rapports trimestriels), de la Presse ou du Gouvernement, la tendance actuelle est à une explosion tant du volume d'information disponible que de sa variété. C'est pourquoi les approches successives de l'extraction d'information ont créé des systèmes de plus en plus modulaires et universels. Par ailleurs, en fonction de la disponibilité de documents annotés, l'extraction peut être supervisée ou non supervisée. À mi-chemin entre ces deux familles d'approches, des approches semi-supervisées peuvent utiliser un nombre réduit d'exemples annotés, un modèle incomplet ou des exemples privés de contextes pour amorcer un système supervisé. Nous nous placerons ici dans le cadre de l'extraction supervisée, présentée à la section 2. Différentes campagnes d'évaluation ont fortement guidé le développement de cette tâche et font l'objet de la section 3. L'extraction supervisée d'informations textuelles faisant appel à des ressources linguistiques et des connaissances fortement dépendantes du domaine, nous analyserons à la section 4 les différentes approches possibles en nous concentrant sur la transition vers les approches neuronales et les représentations distribuées et leur utilité pour le développement de systèmes génériques, donc plus robustes et facilement adaptables à de nouveaux domaines.

2 Présentation des tâches

Ainsi que l'illustre la Figure 1, l'extraction d'information supervisée au sein de sources bruitées et non structurées est une tâche complexe, décomposable en plusieurs sous-tâches généralement traitées séquentiellement. L'extraction d'événements est envisagée comme une tâche de remplissage de formulaire : le type de formulaire correspond à un type d'événement et impose le remplissage d'un nombre variable de champs identifiant les rôles associés à ce type d'événement. Le système extrait alors deux types d'informations : des mentions et des liens entre ces mentions, notamment leur rôle. Cette tâche est donc relativement proche de l'extraction de relations binaires, non présentée ici.

2.1 Extraction de mentions

Reconnaissance d'entités nommées La première étape d'un système d'extraction d'information consiste à identifier dans le texte l'ensemble des entités pouvant remplir un rôle vis-à-vis d'un événement. Une même entité pou-

Ce travail a été partiellement financé par l'ANR dans le cadre du projet ASRAEL (ANR-15-CE23-0018).

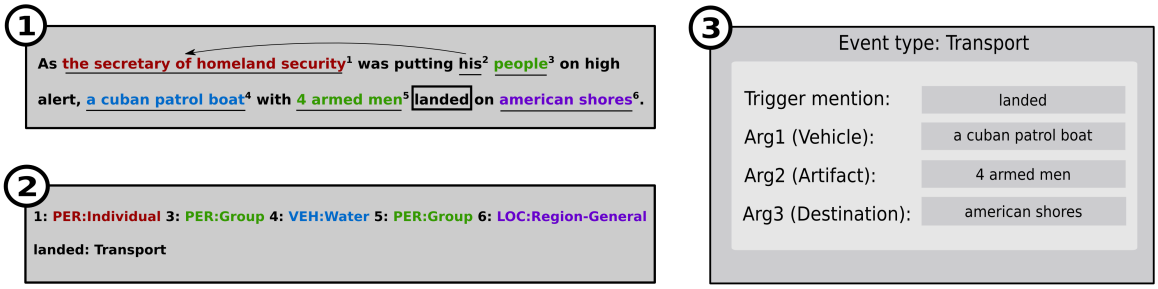

FIGURE 1 – *La reconnaissance d'entités nommées* identifie les différentes mentions d'entités (soulignées dans le cadre (1)) de la phrase et leur type d'entité (cadre (2)). La *résolution de coréférences* permet d'identifier (ici par une flèche) lorsqu'un pronom ou une mention font référence à une même entité. La *détection d'événement* identifie les triggers de la phrase et leur associe un type. Le type du trigger indique le type du formulaire du cadre (3). Les arguments de ce formulaire sont ensuite sélectionnés parmi les entités identifiées précédemment.

vant apparaître plusieurs fois dans un texte (notamment via des pronoms), il s'agit en fait d'extraire des *mentions d'entités* au sein du texte. Trois types d'entités généralement déclinés en sous-types sont généralement distingués : les entités réellement nommées, telles que les noms propres et acronymes de personnes, de lieux ou d'organisations (ENAMEX), les références temporelles telles que les durées ou les dates (TIMEX) et les valeurs numériques telles que les prix ou les pourcentages (NUMEX).

Détection d'événement La majorité des systèmes font l'hypothèse simplificatrice qu'un événement est intégralement défini dans une seule phrase. Ce parti pris est critiquable [28] mais motivé par la plus grande richesse des informations exploitables à l'échelle phrastique. La détection d'événement revient alors à classer des phrases selon un type d'événement prédéfini. La classification d'événement est alors généralement assimilée à la détection de déclencheurs (ou *triggers*) au sein de la phrase. Cette modélisation, introduite par la campagne d'évaluation ACE, est prédominante parmi les approches récentes. Cette simplification n'est pas intrinsèquement nécessaire mais facilite le développement, et notamment l'identification d'événements multiples au sein d'une phrase.

2.2 Extraction de liens entre mentions

Résolution de coréférence La résolution de coréférence, introduite lors de la campagne MUC-6, vise à identifier au sein du texte des mentions faisant référence à une même entité. Cette étape s'appuie principalement sur l'identification de mentions similaires malgré les possibles variations et sur la résolution d'anaphores pronominales pour identifier les correspondances entre mentions d'entité et entre mentions d'entité et pronoms y faisant référence. Elle permet en particulier d'éviter les redondances et de désambiguïser les pronoms pour l'extraction d'événements et de relations. Bien que cette tâche ait initialement été définie pour les entités, la résolution de coréférences entre événements est également courante et permet notamment la consolidation des extractions.

Extraction d'événement Une fois la phrase associée à un type d'événement donné via l'extraction d'un *trigger*, il reste à identifier les entités jouant un rôle dans celle-ci. Cette tâche consiste à prédire, pour chaque entité nommée, son rôle dans l'événement considéré. Dans un cadre supervisé, il est possible de définir un schéma indiquant les types d'entités autorisés pour les différents rôles, ce qui permet de restreindre les possibilités.

2.3 Approches séquentielles et approches jointes

Nous avons présenté séparément les différentes tâches car elles sont généralement traitées de manière séquentielle. Cependant, des interdépendances existent entre ces différentes étapes et peuvent être exploitées. Les approches jointes concernent généralement la prédiction conjointe de triggers et d'arguments. Les phrases suivantes explicitent l'interdépendance de ces tâches :

1. A cameraman died when an American tank **fired** on the Palestine Hotel.
2. He has **fired** his air defense chief.

Ici, le mot "fired" est ambigu mais la présence du mot "tank" correspond probablement au rôle *instrument*, ce qui permet de déduire qu'il s'agit d'un événement de type *Attack*. Dans la deuxième phrase, puisque "Air Defense chief" est un intitulé de poste, l'interprétation de "fired" en tant que trigger de type *End-Position* est évidente.

3 Campagnes d'évaluation

Nous présentons ici plusieurs campagnes d'évaluations qui ont motivé l'apparition et le développement des tâches présentées précédemment.

3.1 Message Understanding Conferences

Les campagnes d'évaluations MUC [1] sont une série de conférences organisées par le DARPA (*Defense Advanced Research Projects Agency*) afin de stimuler la recherche

1. http://itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/proceedings_index.html

en extraction d'information, initialement pour l'analyse de documents militaires. La première campagne d'évaluation MUC (1987) était relativement peu contrainte, sans critères d'évaluation formels. Elle portait sur des documents traitant de repérages et d'interventions en mer. À partir de la deuxième édition (1989), la tâche fut précisément définie comme le remplissage d'un formulaire. Cette tâche s'est complexifiée au fil des éditions pour atteindre 47 champs différents au sein de 11 formulaires (types d'événements) différents pour la cinquième édition. Le principal tournant dans ce programme fut l'introduction de la précision et du rappel comme métriques d'évaluations lors de MUC-3.
Les résultats obtenus lors de la cinquième édition (1993) furent plutôt satisfaisants mais les systèmes étaient fortement spécialisés. Cette spécialisation implique un lourd travail d'adaptation pour chaque nouveau domaine, loin de l'objectif de conception d'un système générique et universel. Pour répondre à ces attentes, les deux dernières éditions de MUC ont identifié des sous-tâches considérées comme fondamentales et pouvant faire l'objet d'évaluations spécifiques : la reconnaissance d'entités nommées et la résolution de coréférences.

3.2 Automatic Content Extraction

Dans la continuité de MUC, les campagnes ACE [2] présentent une tâche de détection et de suivi d'entités incluant la reconnaissance d'entités nommées et la résolution de coréférences. L'édition 2004 introduit la tâche de détection et de caractérisation de relations, qui consiste à extraire des relations étiquetées selon 24 types. ACE présente également une tâche de détection et de caractérisation d'événements couvrant 6 types d'événements et 33 sous-types.
Les documents annotés proviennent de différentes sources : des dépêches d'agence de presse, des bulletins et débats télévisés, des blogs et groupes de discussion en ligne et enfin des échanges téléphoniques. Cette pluralité de sources permit à ACE 2005 de s'imposer comme un cadre de référence pour l'extraction d'événements, et notamment l'étude de l'adaptation au domaine dans ce contexte. ACE 2008 propose deux versions pour les tâches d'extraction d'entités et de relations. La première, à l'échelle locale, correspond aux tâches d'ACE 2004. La seconde version, à l'échelle du corpus, impose donc la résolution de coréférence multi-document, aussi bien pour les entités que pour les relations.

3.3 Autres campagnes

Il existe plusieurs autres campagnes d'évaluation en extraction d'événements ou sur des tâches similaires. La campagne TAC [3] (*Text Analysis Conférence*) propose une tâche de peuplement de base de connaissances comprenant plusieurs sous-tâches d'extraction d'événement, d'*entity linking* (identifier et faire correspondre les entités d'un texte à celles d'une base de connaissances existante et ajouter les entités manquantes à la base), *slot-filling* (compléter certains champs des formulaires de la base à partir des textes). Il existe également plusieurs campagnes d'évaluation en domaine spécialisé, et plus particulièrement dans le domaine biomédical pour permettre d'exploiter les milliers d'articles publiés chaque jour. Les différentes campagnes BioCreatives [4] et I2B2 [5] fournissent ainsi des jeux de données annotés sur l'identification de relations entre traitements et maladies ou entre protéines et gènes par exemple.

4 Méthodes

Bien que les différentes approches d'extraction d'événement présentent des variations importantes, il est possible de les comparer selon plusieurs axes :

- les systèmes ou les différents modules d'un système peuvent être plus ou moins dépendants du domaine concerné ;
- les modèles peuvent s'appuyer sur un degré variable de connaissances linguistiques ;
- la conception des modèles peut nécessiter plus ou moins de données annotées.

Les premiers systèmes [11] d'extraction d'information à partir d'un texte se voulaient universels. Ils ne faisaient pas l'hypothèse de l'existence d'un domaine particulier ou d'un type d'information spécifique à extraire. Ces systèmes visaient à réaliser une analyse complète du document (syntaxique, sémantique et pragmatique) afin de comprendre le texte dans son ensemble. Bien que cette approche soit théoriquement pertinente pour "résoudre" l'extraction d'information indépendamment du domaine ou de la tâche cible, ces systèmes étaient trop complexes à développer et nécessitaient trop de ressources (les développeurs de TACITUS rapportent 36h de calculs pour les 100 messages du jeu de test MUC-3 [12]) et de connaissances à modéliser. De manière générale, la compréhension de texte suppose une analyse en profondeur de l'intégralité des documents. L'implémentation et l'application de ces traitements s'avèrent impossibles, même à l'heure actuelle. En se fixant une tâche moins ambitieuse par l'introduction de la dimension de domaine et d'information structurée et spécifique (entités, attributs, relations et événements), l'extraction d'information se différencie de la compréhension de texte par la profondeur et la couverture de l'analyse linguistique nécessaire au fonctionnement d'un système.

4.1 Approches à bases de connaissances

Utilisation de motifs lexico-syntaxiques Ce changement de paradigme de conception permet l'apparition, dès la fin des années 1980, des premiers systèmes d'extraction d'information. Ce changement annonce aussi la prédominance des approches intra-phrastiques. En effet, puisque l'objectif n'est plus la compréhension globale du texte mais l'extraction ponctuelle d'informations, de nombreux systèmes ne travaillent dans un premier temps qu'au niveau local. Une phase de consolidation est généralement réali-

2. http://www.itl.nist.gov/iad/mig/tests/ace/
3. https://tac.nist.gov/
4. http://www.biocreative.org/
5. https://www.i2b2.org/NLP/

sée à la fin pour fusionner les formulaires construits localement. Ces premiers systèmes (ATRANS [18], SCISOR [26]) tirent leur efficacité de règles et d'expressions régulières définies manuellement. De ce fait, ils nécessitent toujours une conception complexe réalisée spécifiquement par un expert et propre au domaine cible. Ils se caractérisaient par un aspect rigide et monolithique et étaient difficilement adaptables à d'autres langues ou objectifs.

Le système FASTUS [12] popularise sur la campagne MUC-3 l'utilisation d'automates à états finis en cascades et plus généralement l'approche séquentielle. Ce système réalise ainsi séquentiellement 5 étapes de reconnaissance de motifs et de *chunking*, la sortie de chaque module étant l'entrée du module suivant. L'intérêt de cette décomposition est l'apparition de la modularité au sein du système. Celle-ci permet une modification plus aisée qui facilite le développement et l'adaptation du système. De plus, les trois premières étapes opèrent au niveau linguistique et sont très peu dépendantes du domaine. De ce fait, l'adaptation au domaine ne concerne que les 2 dernières étapes.

Extraction de motifs lexico-syntaxiques Si les systèmes à base d'automates en cascade marquent un tournant au regard de la lourdeur de la tâche de conception d'un système, cette conception est toujours manuelle et nécessite l'intervention de connaissances expertes à la fois sur le système et sur le domaine. Un nouveau changement de paradigme intervient avec l'utilisation de méthodes d'extraction de motifs. Le travail nécessaire pour l'adaptation à un nouveau domaine est ainsi passé de la conception des règles à l'annotation d'un corpus.

Ces méthodes (RAPIER [21], Autoslog [27]) utilisent différentes représentations des exemples telles que des sacs de mots, l'étiquetage en parties du discours ou des arbres syntaxiques. [10] propose une approche séquentielle d'identification des mentions d'événements puis des arguments pour finir par la classification du type d'événements. Ce système s'appuie sur l'utilisation de structures syntaxiques et de classifieurs séquentiels et constitue donc aussi un précurseur des approches à base de classifieurs.

4.2 Apprentissage de classifieurs

À la différence des systèmes précédents, les systèmes utilisant des classifieurs considèrent la tâche d'extraction d'événement comme une tâche de classification de séquence. Un texte est un ensemble de phrases traitées comme des séquences de *tokens*. La détection d'événement consiste alors à appliquer à chaque élément de la séquence un classifieur entraîné à détecter les *triggers* et leur type, séquentiellement ou de manière jointe. Il en va de même pour les arguments, généralement prédits parmi les entités nommées détectées en amont.

On dénote au sein de cette famille d'approches deux tendances. La majorité des études utilisent des approches séquentielles en traitant d'abord l'identification puis la classification des *triggers* puis des arguments ([1, 6, 10]). Mais certaines études utilisent également des approches jointes [4, 15]. Ces méthodes tentent de réduire le problème de propagation des erreurs symptomatique des approches séquentielles. De plus, elles peuvent ainsi tenir compte de l'interdépendance entre arguments et *triggers* ou entre détection et caractérisation des *triggers* ou des arguments. Néanmoins, ces approches se rejoignent sur les types de classifieurs et de représentations choisis. Ces approches ont évolué parallèlement à celles concernant l'extraction de relations. C'est pourquoi nous citerons ici indifféremment des études portant sur les deux tâches. Les classifieurs sont le plus souvent des machines à vecteurs de support [33, 17, 13] ou des classifieurs de type maximum d'entropie [29, 23]. L'efficacité de ces approches étant particulièrement dépendante de la qualité des représentations choisies, la création de représentations adaptées est essentielle. Les approches à bases de classifieurs n'ont ainsi supprimé l'effort d'élaboration de règles que pour le remplacer par un effort d'ingénierie des représentations aussi conséquent.

Il apparaît cependant qu'une fois des représentations efficaces obtenues, celles-ci s'avèrent assez génériques pour être transposées dans des domaines proches. [33] introduit pour l'extraction de relations la plupart des traits (*features*) utilisés dans l'état de l'art. Ces représentations sont produites à plusieurs niveaux : au niveau lexical (sac de mots et tête de mention pour chaque mention, premiers et deuxièmes mots des contextes gauche, milieu, et droit), syntaxique (chemins dans l'arbre syntaxique complet entre les deux mentions, *chunking* puis extraction des têtes des groupes nominaux) et sémantique (utilisation des types d'entités ACE et de WordNet [20]). [29] reprend ces représentations et complète la représentation lexicale par l'utilisation de bigrammes des mots du contexte central. D'autres informations sémantiques sont proposées, telles que l'utilisation de synonymes de WordNet [15] ou d'hyperonymes de Framenet [16]. Il est à noter que le niveau de granularité maximum de ces représentations est généralement le mot bien que des approches descendent au niveau des morphèmes pour l'extraction d'information en chinois [6]. La représentation des mots est généralement de type local ou *one-hot*, c'est-à-dire par un vecteur binaire de taille N où N est la taille du vocabulaire et dont seule la dimension correspondant au mot est active. Cette représentation symbolique pose deux problèmes [30] : d'une part elle ne permet pas de capturer convenablement la sémantique du mot (ce à quoi tentent de pallier les approches par morphèmes) ; d'autre part, elle est particulièrement parcimonieuse et sujette au fléau de la dimension. Enfin, si les vocabulaires cible et source sont différents, le système n'aura aucune information sur les mots nouveaux.

4.3 Plongement lexical

Le plongement lexical (ou *word embeddings*) est une représentation distribuée des mots permettant de répondre aux deux problèmes soulevés à la section précédente. Dans une représentation locale (ou symbolique), un élément est associé à une représentation unique (un indice). Au contraire,

dans une représentation distribuée, un élément est décrit par plusieurs indices et un indice est utilisé pour décrire plusieurs éléments. Si la représentation locale est plus facile à comprendre et à produire par un humain, elle ne permet pas de capturer la proximité sémantique entre les éléments ou d'isoler différentes propriétés sous-jacentes.

Pour le plongement lexical, l'hypothèse distributionnelle ("You shall know a word by the company it keeps !" [9]) postule que des mots apparaissant dans des contextes similaires ont des sens similaires. On peut apprendre ces représentations sur de grands corpus de textes non annotés. Ceci confère un autre avantage à ces représentations : elles permettent d'assurer une meilleure robustesse aux systèmes d'extraction d'information, même pour des mots non présents dans le jeu d'apprentissage. Plusieurs approches successives ont proposé des représentations distribuées, notamment l'analyse sémantique latente [8] et le clustering de brown [3]. Il apparaît ensuite une série de représentations [2, 7, 19] extraites de réseaux de neurones entraînés à prédire un mot à partir de son contexte ou inversement. Leur utilisation en extraction de relations commence avec [29] qui exploite des *clusters* de Brown pour augmenter les descripteurs des têtes de mentions.

4.4 Architectures neuronales

Peu après l'introduction de ces représentations distribuées et en réponse aux propagations d'erreurs et aux difficultés d'adaptation inhérentes aux systèmes séquentiels, de nombreuses études ont commencé à appliquer des approches neuronales à différentes tâches de l'extraction d'information. Elles mettent en avant la suppression des prétraitements à l'origine de ces problèmes. Ces approches neuronales utilisent différentes architectures de réseaux de neurones popularisées dans les communautés de reconnaissance de la parole et de vision par ordinateur. Plusieurs systèmes d'extraction de relations [14, 32] proposent des systèmes extrayant une représentation de la phrase à l'aide d'un réseau convolutif (CNN). Ces systèmes ne fournissent en entrée que les *word embeddings* des *tokens* de la phrase, généralement ceux de [19]. Ces *word embeddings* sont modifiés durant l'apprentissage, ce qui permet de les adapter au domaine cible. L'opération de *pooling* du réseau convolutif ne conservant que l'information prédominante d'une phrase, [5] propose une variante du CNN utilisant le *dynamic multipooling* (ou *piece-wise CNN* [31]), c'est-à-dire l'extraction de représentations pour plusieurs parties de la phrase, ce qui permet notamment de mieux gérer les phrases contenant de multiples événements. Par ailleurs, l'application de convolutions est surtout adaptée à la détection de motifs locaux et consécutifs. Des systèmes utilisent au contraire des réseaux récurrents pour construire des représentations pouvant tirer profit de motifs non consécutifs à l'échelle de la phrase. De manière générale, un des intérêts de ces méthodes neuronales est d'obtenir une abstraction de la représentation d'entrée, de telle sorte que la représentation en sortie soit relativement invariante à de faibles changements locaux, augmentant ainsi la robustesse de la représentation finale.

TABLE 1 – Résultats de différents systèmes de détection d'événements (F1-mesure pour développement et test)

Méthodes	**dév.**	**test**
Meilleure méthode non neuronale : prédiction structurée [15] avec descripteurs locaux et globaux	67,9	67,5
CNN avec *embeddings* (mots) [24]	14,0	–
CNN avec *embeddings* (mots + positions) [24]	68,5	67,6
CNN avec *embeddings* (mots + positions + entités) [24]	70,7	69,0

4.5 Généralisation des représentations distribuées

Les différents systèmes ayant redéfini l'état de l'art à l'aide des approches neuronales ont tous mis en avant l'objectif de réduction des prétraitements nécessaires, notamment syntaxiques et sémantiques, pour se concentrer sur les représentations lexicales distribuées. On peut cependant rapidement voir la réintroduction d'informations syntaxiques puis sémantiques. Dans un premier temps, pour l'extraction de relations, la représentation en entrée du réseau est augmentée d'un vecteur codant la distance aux deux entités concernées (*position feature*) et qui permet, à titre indicatif, de gagner dans [32] plus de 9 points en F1-mesure pour l'extraction de relations et jusqu'à 54 points en extraction d'événements dans [24] (cf. Table 1). La majorité des systèmes augmente également la représentation de chaque *token* par un *embedding* de son type d'entité (*entity feature*, +2 points) [24], ou d'informations syntaxiques extraites d'arbres syntaxiques ou d'un *chunker* [22, 25]. Ces systèmes montrent bien que ces différentes informations linguistiques initialement supprimées des approches neuronales sont importantes pour le développement de systèmes performants. L'utilisation de représentations distribuées pour représenter ces informations confère plusieurs avantages déjà identifiés pour les représentations distribuées de mots. D'une part, ces représentations sont plus à même d'isoler les facteurs sous-jacents à ces différentes informations. D'autre part, ces représentations dans un espace dense sont plus expressives et permettent au réseau de tenir compte de la similarité entre différentes modalités. Enfin, puisque ces représentations sont apprises et transférables, elles permettent une meilleure adaptation à de nouveaux domaines et offrent la possibilité d'un apprentissage semi-supervisé. Ainsi, de la même manière que le pré-entraînement de *word embeddings* sur une base annexe offre une meilleure généralisation, cet apprentissage semi-supervisé est possible pour les autres types de représentations distribuées.

Conclusion et perspectives

Les premiers systèmes d'extraction d'information nécessitaient l'élaboration manuelle de motifs lexico-syntaxiques finement adaptés à une tâche et un domaine donnés. Ceci

les rendaient coûteux et complexes à développer et particulièrement sensibles à un changement de domaine ou un décalage entre les jeux d'apprentissage et les jeux de test. Les approches suivantes utilisant des documents annotés nécessitaient toujours un temps de développement long, tout en étant plus génériques et modulables. L'introduction des représentations distribuées et des réseaux de neurones a ensuite grandement amélioré l'état de l'art. La synergie de ces différentes représentations permet de combler les faiblesses des précédents algorithmes sur plusieurs points. D'une part, elles permettent de réaliser un apprentissage semi-supervisé grâce à des bases annexes, notamment pour les *word embeddings* et d'autre part, elles offrent une meilleure expressivité et une meilleure interprétabilité par les classifieurs que les représentations symboliques. Cependant, l'abandon de ces dernières se fait au détriment de l'interprétabilité humaine du système, constat déjà réalisé pour la reconnaissance d'objets en vision. À l'avenir, ces représentations distribuées pourraient notamment être appliquées à l'extraction d'information ouverte en transférant des représentations apprises de manière semi-supervisée à cette tâche d'extraction peu ou pas supervisée.

Références

[1] David AHN. « The Stages of Event Extraction ». *Workshop on Annotating and Reasoning about Time and Events*.

[2] Yoshua BENGIO et al. « A Neural Probabilistic Language Model ». *Journal of Machine Learning Research* 3 (2003).

[3] Peter F. BROWN et al. « Class-Based n-Gram Models of Natural Language ». *Computational linguistics* 18.4 (1992).

[4] Chen CHEN et Vincent NG. « Joint Modeling for Chinese Event Extraction with Rich Linguistic Features ». *COLING*. 2012.

[5] Yubo CHEN et al. « Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks ». *ACL-IJCNLP*. 2015.

[6] Zheng CHEN et Heng JI. « Language Specific Issue and Feature Exploration in Chinese Event Extraction ». *NAACL-HLT*. 2009.

[7] Ronan COLLOBERT et Jason WESTON. « A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning ». *ICML*. 2008.

[8] S. T. DUMAIS et al. « Using latent semantic analysis to improve access to textual information ». *SIGCHI*. 1988.

[9] John R FIRTH. « A Synopsis of Linguistic Theory, 1930-1955 ». *Studies in Linguistic Analysis* (1957).

[10] Ralph GRISHMAN, David WESTBROOK et Adam MEYERS. « NYU's English ACE 2005 System Description ». *ACE*. 2005.

[11] Jerry R. HOBBS. « Overview of the TACITUS project ». *The finite string newsletter* (1986).

[12] Jerry R. HOBBS et al. « FASTUS : A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text ». *Finite-state language processing* (1997).

[13] Yu HONG et al. « Using Cross-Entity Inference to Improve Event Extraction ». *ACL-HLT*. 2011.

[14] Yoon KIM. « Convolutional Neural Networks for Sentence Classification ». *EMNLP*. 2014.

[15] Qi LI, Heng JI et Liang HUANG. « Joint Event Extraction via Structured Prediction with Global Features. » *ACL*. 2013.

[16] Qi LI et al. « Constructing Information Networks Using One Single Model. » *EMNLP*. 2014.

[17] Shasha LIAO et Ralph GRISHMAN. « Using Document Level Cross-Event Inference to Improve Event Extraction ». *ACL*. 2010.

[18] Steven L. LYTINEN et Anatole GERSHMAN. « ATRANS Automatic Processing of Money Transfer Messages. » *AAAI*. 1986.

[19] Tomas MIKOLOV et al. « Distributed Representations of Words and Phrases and Their Compositionality ». *Advances in Neural Information Processing Systems*. 2013.

[20] George A. MILLER. « WordNet : A Lexical Database for English ». *Communications of the ACM* 38.11 (1995).

[21] Raymond J. MOONEY et Mary E. CALIFF. « Relational Learning of Pattern-Match Rules for Information Extraction ». *AAAI*. 1999.

[22] Thien H. NGUYEN, Kyunghyun CHO et Ralph GRISHMAN. « Joint Event Extraction via Recurrent Neural Networks ». *NAACL-HLT*. 2016.

[23] Thien H. NGUYEN et Ralph GRISHMAN. « Employing Word Representations and Regularization for Domain Adaptation of Relation Extraction. » *ACL*. 2014.

[24] Thien H. NGUYEN et Ralph GRISHMAN. « Event Detection and Domain Adaptation with Convolutional Neural Networks ». *ACL-IJCNLP*. 2015.

[25] Thien Huu NGUYEN et Ralph GRISHMAN. « Combining Neural Networks and Log-Linear Models to Improve Relation Extraction ». *IJCAI Workshop on Deep Learning for Artificial Intelligence*. 2016.

[26] Lisa F. RAU. « Conceptual information extraction from financial news ». *HICSS*. 1988.

[27] Ellen RILOFF. « Automatically Constructing a Dictionary for Information Extraction Tasks ». *AAAI*. 1993.

[28] Mark STEVENSON. « Fact Distribution in Information Extraction ». *Language Resources and Evaluation* 40 (2006).

[29] Ang SUN, Ralph GRISHMAN et Satoshi SEKINE. « Semi-Supervised Relation Extraction with Large-Scale Word Clustering ». *ACL-HLT*. 2011.

[30] Joseph TURIAN, Lev RATINOV et Yoshua BENGIO. « Word Representations : A Simple and General Method for Semi-Supervised Learning ». *ACL*. 2010.

[31] Daojian ZENG et al. « Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. » *EMNLP*. 2015.

[32] Daojian ZENG et al. « Relation Classification via Convolutional Deep Neural Network. » *COLING*. 2014.

[33] GuoDong ZHOU et al. « Exploring Various Knowledge in Relation Extraction ». *ACL*. 2005.