
Préface

Sophia ANANIADOU
University of Manchester
sophia.ananiadou@manchester.ac.uk

Nathalie FRIBURGER
LI, Université de Tours
nathalie.friburger@univ-tours.fr

Sophie ROSSET
LIMSI-CNRS, Orsay
sophie.rosset@limsi.fr

Les entités nommées constituent un champ de recherche très actif et sont considérées comme un point central dans de nombreuses applications (par exemple, la recherche d'information précise) mettant en jeu des notions comme la compréhension, la recherche sémantique, etc.

La notion d'entités nommées (EN) couvre non seulement les noms propres mais aussi des entités plus complexes comme les expressions multimots. Elle recouvre classiquement des noms désignant des personnes, des lieux ou des organisations mais peut aussi se rapporter à des notions spécialisées comme les maladies, les noms de gènes, les composés chimiques, etc. Ces entités nommées sont donc, en général, typées selon des taxonomies plus ou moins vastes et fortement dépendantes du domaine d'application (biologie, médecine, etc.) ou des besoins considérés (recherche d'information, résumé automatique ou encore traduction automatique).

La première tâche de détection d'entités nommées a été proposée en 1995 par la sixième édition de la campagne d'évaluation MUC (*Message Understanding Conference*) et a été dès le départ considérée comme une tâche élémentaire pour la compréhension de textes. Dès le début, aux noms propres, ont été ajoutées les expressions numériques représentant les expressions temporelles et les montants monétaires. Ces entités nommées et numériques constituent ainsi une brique élémentaire sur laquelle l'analyse de document repose. La définition même d'une entité nommée et de la réalité linguistique qu'elle recouvre a beaucoup évolué depuis cette première proposition. Les travaux concernant les EN se sont structurés notamment autour de campagnes d'évaluation qui ont contribué à l'évolution de la définition des EN, que ce soit dans le domaine dit général (MUC, ACE, HAREM, EVALITA, ESTER, ETAPE, etc.) ou de spécialité (BioNLP, BioCreative, I2B2, etc.). Actuellement, les EN ne couvrent plus

seulement des noms propres mais aussi des groupes nominaux et parfois même adjectivaux dès lors qu'ils renvoient à un objet du discours identifiable. Les catégories ont également été étendues et surtout affinées, allant jusqu'à distinguer 300 types différents. Enfin, des propositions ont été faites pour structurer les éléments qui composent une entité et les typer.

Ces différentes définitions sont explorées par les contributions de ce numéro spécial. Si, lors des premiers travaux, la détection des EN était envisagée dans des documents textuels, elle a rapidement porté aussi sur des données ayant subi un traitement préalable comme la reconnaissance optique de caractères (pour des documents anciens) ou la reconnaissance automatique de la parole (pour des données orales). La détection des relations entre EN est en quelque sorte l'étape suivant celle de détection des EN. Quelles relations entretiennent-elles entre elles ? Comment trouver ces relations ? Enfin, les EN, parce qu'elles sont des éléments à la base de l'information contenue dans un document, peuvent contribuer à des applications TAL comme la recherche d'information, le résumé automatique ou la traduction automatique entre autres.

Pour ce numéro spécial, douze soumissions ont été reçues (dont cinq en anglais) couvrant un large spectre de thèmes. Au final, quatre articles ont été retenus portant sur la détection d'entités nommées structurées (Nouvel *et al.*), la détection d'entités nommées dans des données de parole transcrites automatiquement (Hatmi *et al.*), la détection de relations (Wang *et al.*) et les entités nommées pour la traduction automatique (Gahbiche-Braham *et al.*). Les approches présentées sont tout autant variées, allant de la fouille de données à l'apprentissage non supervisé. Comme on peut s'y attendre, ces articles abordent différents domaines complémentaires en relation avec les entités nommées.

Dans cette introduction, nous souhaitons présenter d'une part les aspects de l'état de l'art qui sont abordés par ces différents articles, d'autre part un résumé thématique des travaux soumis.

Ces articles proposent des états de l'art plus ou moins fournis dans différents domaines en lien avec les entités nommées : la REN (reconnaissance d'entités nommées), les corpus, les relations entre entités nommées et la traduction des entités nommées. Trois des articles présentent la reconnaissance d'entités nommées (REN) sous trois angles différents, le quatrième aborde la notion de relations entre EN.

Ainsi, Nouvel *et al.* comparent les systèmes de reconnaissance d'entités nommées orientés connaissances à base de règles (notamment avec des transducteurs) avec l'approche orientée données qui utilise des algorithmes d'apprentissage (CRF, par exemple). Cet article présente aussi des travaux proches de la REN utilisant des techniques de fouille de textes, techniques qui, habituellement, ne sont pas utilisées pour la REN. Au-delà des techniques de la REN, le type de données et de modalités (écrit/oral) dans lesquelles on cherche à détecter les entités nommées nécessite d'être pris en compte. Hatmi *et al.* présentent un état des lieux de la REN dans le cadre particulier de la parole ainsi que les problèmes liés à cette modalité (ambiguïtés graphiques et de segmentations dues à l'absence de capitalisation et au manque de ponctuation,

disfluences, bruit dans les transcriptions). Un dernier aspect à considérer est la langue. Gahbiche-Braham *et al.* présentent leurs travaux sur la détection des EN en langue arabe et un état de l'art sur les travaux récents de REN fondés sur des méthodes par règles ou apprentissage dans cette langue.

Afin d'évaluer les résultats des systèmes de REN, il est nécessaire d'avoir des corpus annotés en EN. Deux des articles de ce numéro donnent de nombreuses informations sur les corpus (en français) et les campagnes d'évaluation liés aux entités nommées. Hatmi *et al.* présentent une description du corpus utilisé lors de la campagne d'évaluation ESTER2, ses différentes parties (entraînement, développement et test) et son guide d'annotation. Cet article expose ensuite les mesures de performance utilisées pour évaluer les résultats des systèmes. Comme les précédents, Nouvel *et al.* décrivent également le corpus ESTER2. Ils s'attachent aux caractéristiques de ce corpus en termes de répartition des différents types d'EN et discutent des problèmes liés aux erreurs d'annotation. Les auteurs décrivent ensuite le corpus de la campagne ETAPE, utilisant l'annotation Quaero, plus fine que celle d'ESTER2, donc potentiellement plus difficile pour les systèmes.

Un thème important abordé par (Wang *et al.*) est l'extraction de relations entre EN. Ils nous présentent un état de l'art de ce thème et nous rappellent que les relations entre EN sont souvent représentées par des relations binaires ou adoptent des patrons. Les relations candidates sont le plus souvent sélectionnées par *clustering* ou l'usage d'algorithmes d'ordonnancement. Ces systèmes reposent sur des modèles d'apprentissage statistiques supervisés, ou bien sur des patrons lexico-syntaxiques ou règles.

Enfin, un dernier point porte sur la traduction des EN dont Gahbiche-Braham *et al.* nous présentent un état de l'art. Différentes méthodes permettent de traduire les EN d'une langue (ici, l'arabe) vers une autre. L'arabe ayant un alphabet particulier, de nombreux travaux s'appuient sur la translittération des mots inconnus de l'arabe en alphabet latin. La forme translittérée est ensuite recherchée dans des dictionnaires bilingues ou monolingues avec l'aide de méthodes statistiques. Cette approche ne convient pas à toutes les EN et des alternatives sont aussi présentées dans cet article (corpus parallèles, transducteurs, etc.).

Les articles de ce numéro présentent des travaux sur la reconnaissance d'entités nommées mais aussi sur les relations entre entités nommées et leur traduction. Parmi ces quatre articles, deux concernent directement la reconnaissance d'entités nommées avec, pour l'un, des expériences d'hybridation de systèmes de REN entre eux et pour l'autre, le couplage d'un système de REN et de reconnaissance automatique de la parole (RAP).

Ainsi, Nouvel *et al.* s'attachent à la détection d'EN structurées et compositionnelles. Ils proposent une approche originale qui combine la fouille de règles d'association multiniveaux, proposant des balises à insérer, ainsi qu'un classifieur fondé sur une approche supervisée. Pour cela, ils présentent leurs expériences d'hybridation entre un système d'extraction de connaissances (mXS) et un système à base de transducteurs (CasEN) dans le contexte de la participation aux campagnes ESTER2 et ETAPE. Ils décrivent la préparation des données en utilisant des ressources morphosyntaxiques et lexicales pour annoter et enrichir les textes. La fouille de ces données enrichies permet

de trouver de très nombreux motifs (règles) sur lesquels sont appliqués des principes de généralisation, ce qui permet de procéder à une extraction de règles d'annotation partielles selon différents seuils de fréquence et de confiance. Ces règles permettent ensuite d'annoter automatiquement des textes en EN. Les performances, très satisfaisantes, sont évaluées sur les corpus ESTER2 et ETAPE et comparées à celles d'autres systèmes.

Hatmi *et al.* s'intéressent plus particulièrement à des systèmes de REN lorsqu'ils sont appliqués sur des corpus oraux transcrits par un système de reconnaissance automatique de la parole ; en effet, dans ce cas, les erreurs restantes rendent plus complexe la tâche de reconnaissance des EN. Ils proposent donc d'intégrer la reconnaissance des EN dans le processus de décodage de la parole : la tâche de REN (réalisée par le système LIA_NE) est ainsi couplée au système de transcription dans le système SRAP_REN. Il s'agit d'annoter tout d'abord le corpus d'apprentissage en EN ainsi que le vocabulaire du système de transcription puis d'adapter les modèles de langage. Ces modèles sont réapppris sur le corpus étiqueté en EN pour permettre au système de transcription de générer des sorties annotées en EN. Cette méthode a permis une amélioration de la qualité de la reconnaissance des entités.

Les systèmes de REN présentés dans ces deux articles ont été testés et évalués sur des données orales transcrites, données sur lesquelles le défi de la REN est plus difficile à relever.

Les deux articles suivants s'attachent à des problèmes qui vont au-delà de la simple détection des EN.

Wang *et al.* se focalisent sur la détection de relations entre entités nommées, une tâche qui se situe entre la détection des EN et le remplissage de formulaire tel qu'envisagé par les premières campagnes MUC puis dans le cadre de la campagne KBP (*Knowledge Base Population*). Ils présentent une méthode pour extraire des relations entre entités nommées et les regrouper en *clusters*. L'extraction des relations se fonde sur l'identification de couples d'entités nommées dans une même phrase puis sur des heuristiques de filtrage afin de supprimer les relations sans intérêt. Ensuite le regroupement des relations (*clustering*) se fait à deux niveaux par des mots-clés représentatifs et par la similarité sémantique entre mots et relations, regroupés par thématiques des relations. Les résultats des différents *clusterings* sont évalués et permettent de conclure que cette méthode a la capacité d'extraire et regrouper les relations avec une bonne précision.

Enfin, Gahbiche-Braham *et al.* présentent leur système de détection des EN en arabe (NERAr) : le but de cette détection d'EN en arabe est d'améliorer leur système de traduction automatique (arabe vers français) avec des connaissances externes (dictionnaires). Les auteurs utilisent un corpus d'apprentissage de données de l'AFP et présentent différentes versions du modèle de détection des EN avec des jeux de traits portant sur la ponctuation et les nombres, les N-grammes de mots, les préfixes et suffixes, les parties du discours (étiquettes POS), etc. La détection des EN sert de prétraitement à la traduction : une fois les EN détectées, on les traduit grâce à des dictionnaires puis on passe à la phase de traduction complète du texte.

Ce numéro spécial de la revue TAL offre un échantillon des travaux actuels portant sur la détection des entités nommées, les relations entre entités nommées et la traduction. On remarque que les domaines de recherche actuels en rapport avec les entités nommées comme les domaines de spécialité, les questions de l'évaluation et des métriques, la définition même des entités nommées, la normalisation des données, la résolution des coréférences, etc. manquent à ce numéro.

Remerciements

Nous souhaitons exprimer notre gratitude envers l'Institut des sciences humaines et sociales du CNRS qui soutient la revue TAL, le comité scientifique de la revue TAL et les membres du comité scientifique spécifique de ce numéro spécial :

Maud Ehrmann, European Commission, JRC, Italy
Olivier Galibert, LNE, France
Natalia Grabar, STL, Université de Lille 1 et 3, France
Kais Haddar, University of Sfax, Tunisie
Thierry Hamon, LIM&Bio, Paris 13, France
Valia Kordoni, Humboldt-Universität, Berlin, Germany
Anna Korhonen, University of Cambridge, UK
Ioannis Korkontzelos, University of Manchester, UK
Anne-Laure Ligozat, LIMSI, France
Bernardo Magnini, FBK, HLT, Italy
Makoto Miwa, University of Manchester, UK
Claire Nedellec, MIG, INRA, France
Aurélie Névéol, LIMSI, France
Noaoaki Okazaki, Tohoku University, Japan
Christian Raymond, IRISA, France
Fabio Rinaldi, University of Zurich, Swiss
Patrick Ruch, University of Geneva, Swiss
Benoît Sagot, ALPAGE, INRIA, France
Agata Savary, LI, Université François Rabelais Tours, France
Satoshi Sekine, NYU, USA
Jian Su, A-STAR, Singapore
Junichi Tsujii, Microsoft Research Asia, China
Patrick Watrin, UCL, CENTAL, Belgique
Fabio Zanzotto, Tor Vergata, university of Rome, Italy