http://dx.doi.org/10.18778/8142-075-4.06

LA RECONNAISSANCE D'ENTITÉS NOMMÉES ET LA GRANULARITÉ DES RESSOURCES DICTIONNAIRIQUES

Krzysztof Bogacki
Université de Varsovie

Abstract. In this article we briefly report on different architectures of the computer systems dedicated to the identification of named entities in texts, such as person name, location, organization and date. These are able to identify named entities with a precision and a recall rate exceeding 0.9. We then discuss two types of named entities designating people and show the importance of the granularity of semantic descriptions associated with each entry in very large coverage electronic dictionaries contained in the named entity recognition system.

Introduction

On a maintes fois décrit les difficultés qu'on retrouve en voulant définir le concept de mot de façon univoque sans laisser de résidu. Une des réactions à cet état de choses a été de proposer le terme d'entité nommée. Il est né dans le domaine de la linguistique computationnelle dans les années 80 du siècle dernier en réponse à un besoin pratique – nécessité de reconnaître un certain type d'information contenue dans des textes. Il avait reçu une définition[1] qui, au départ, englobait l'ensemble des noms de personnes, d'organisations, d'entreprises et de lieux (ENAMEX), d'expressions temporelles (TIMEX) avec trois sous-classes (dates, heures et périodes de temps) et d'autres données chiffrées (NUMEX : montants financiers, pourcentages, etc.). D'autres divisions ont été proposées ensuite déplaçant les limites initialement fixées. En effet, assez rapidement on a allongé cette liste en

[1] Cf. les 7 campagnes d'évaluation des systèmes informatiques MUC (= Message Understanding Conferences) servant à extraire des données précises des textes. Cf. M. Ehrmann (2008 : 21).

y ajoutant des entités désignant les éléments de base typiques pour un domaine donné, p. ex. noms d'animaux dans une étude de zoologie ou noms de molécules chimiques dans un traité de pharmacologie, etc. L'utilité de ces séquences s'est avérée primordiale surtout dans la fouille de données et dans d'autres applications du domaine d'extraction d'information (p. ex. pour l'indexation des documents par les moteurs de recherche) en suscitant en même temps un certain nombre de réflexions théoriques sur les rapports entre cette notion qui venait d'émerger grâce aux informaticiens et d'autres concepts ayant leurs lettres d'ancienneté en philosophie et en linguistique.

Sur le plan théorique, ce qui frappe, c'est qu'on donne souvent une définition énumérative du concept d'entité nommée recouvrant une réalité hétérogène, extrêmement flexible car pouvant s'adapter à diverses situations. D'un point de vue référentiel, les entités nommées en tant que noms de fragments de la réalité extratextuelle sont proches des noms propres (cf. Kripke, 1982) sans se confondre avec eux. En effet, *le président de la République Française* identique référentiellement à *Emmanuel Macron* n'est pas un nom propre. D'autre part, la monoréférentialité ne garantit pas qu'une expression candidate soit classée entité nommée (cf. le pronom *je* discursivement monoréférentiel). Il serait erroné de vouloir dresser une liste exhaustive d'entités nommées qui ne cesse d'augmenter au rythme de publication de textes. Toutes les tentatives de formuler une définition classique descriptive se sont soldées par des échecs[2]. Seule une définition stipulative serait concevable : elle énumérerait les types de syntagmes ou de mots monoréférentiels considérés comme entités nommées dans les circonstances précises.

Dans la suite de cet article que nous offrons en hommage à Alicja Kacprzak, collègue et amie de longue date, après avoir signalé rapidement différentes architectures de systèmes informatiques dédiés au repérage des entités nommées dans les textes, nous allons réfléchir à quelques éléments strictement linguistiques impliqués par ces outils informatiques.

1. Types de systèmes d'identification et d'extraction d'entités nommées

Plusieurs extracteurs ont été développés pour différentes langues. Les résultats atteints sont impressionnants : la fiabilité de solutions proposées dépasse parfois 0,90 en termes de rappel et de précision. Différentes techniques sont mises en œuvre.

[2] Cf. Ehrmann (*op. cit.*, p. 255-258) pour la liste des 13 formules définitoires existantes.

D'un côté, mentionnons l'approche symbolique qui est sans doute la plus utilisée. Elle fait appel à des règles écrites manuellement par des experts du domaine. Des connaissances linguistiques sont nécessaires pour établir une liste de règles d'annotation qui tiennent compte aussi bien des constituants de l'entité nommée que de leur contexte. La procédure de la reconnaissance comporte trois étapes.

Tout d'abord, en s'aidant d'outils informatiques spécialisés, on extrait du texte les caractéristiques des termes grâce à leurs propriétés morphosyntaxiques de surface, ce qui conduit à un étiquetage fonctionnel. Les balises placées dans le texte (p. ex. signalant le début (<personne>) et la fin (</personne>)) enrichissent la représentation des mots et seront exploitées au moment de l'application des règles symboliques.

On compare ensuite le texte à tour de rôle avec deux types de dictionnaires inclus dans le logiciel. Tout d'abord avec ceux qui comportent les unités du lexique général. Cette consultation est faite dans l'espoir d'identifier les « déclencheurs » ou « termes saillants » qui signalent les entités nommées. Le système dispose aussi des listes de noms propres de toute sorte : patronymes, prénoms, sobriquets, toponymes, noms d'organisations, sigles, etc. et des étiquettes sémantiques en relation avec les entités nommées.

La dernière étape consiste à désambiguïser les entités qui restent confuses soit pour des raisons sémantiques, soit structurelles.

Le deuxième type d'approche est basé sur la statistique et connaît trois variantes : approche par apprentissage supervisé, semi-supervisé et non supervisé. Les systèmes de ce type sont souvent utilisés dans le cas des textes bruités (p. ex. dans le cas des transcriptions automatiques des émissions radio). Les deux premiers exploitent un ensemble d'exemples annotés et manipulent de grandes masses de textes en essayant d'en extraire des règles d'annotation. Dans un premier temps, les systèmes sont entraînés à exploiter les traits singularisant les entités nommées avant de généraliser le processus sur de nouveaux documents. Dans les systèmes basés sur l'apprentissage non supervisé, on traite des exemples non annotés. Il existe aussi des méthodes hybrides combinant les deux démarches : linguistique et statistique.

Quel que soit le type d'approche, on doit s'assurer de la reconnaissance morphologique des mots dans le texte analysé. Pour certaines langues avec un système morphologique compliqué, p. ex. le polonais, la confection d'un outil permettant l'identification des formes fléchies s'avère un premier obstacle ardu exigeant un effort considérable[3].

[3] Les dimensions pour le polonais sont de l'ordre de 450.000 mots vedettes et le nombre d'unités fléchies dépasse 22 millions de formes.

2. Entité nommée et les ressources dictionnairiques

Il est évident qu'on cherche à doter les vedettes de la liste des noms propres du plus grand nombre d'étiquettes, ce qui permettra de faire face à des requêtes exceptionnelles qui s'écartent de l'association standard d'un nom comme *Donald Trump* avec l'étiquette *président des États-Unis* ou *Émile Zola* avec *écrivain.* On note en effet des situations où les noms de personnes sont employés pour les ouvrages dont ces personnes sont les auteurs : *Luc est un grand spécialiste qui connaît tout sur la V^e^ République. Figure-toi qu'il a lu* ***tout De Gaulle****.* Il arrive qu'un même nom propre renvoie, selon le contexte, à une institution, à une communauté d'individus ou encore à un bâtiment :

> Le journal télévisé a été diffusé hier en direct de **la Bourse de Paris**.
> **La Bourse de Paris** a inauguré hier sa session avec un quart d'heures de retard à cause d'une panne d'électricité.
> **La Bourse de Paris** va fêter son centenaire l'année prochaine.
> **La Bourse de Paris** a réagi mal au dernier attentat terroriste dans le métro parisien.

L'examen des textes servant de base pour la reconnaissance des entités nommées fait voir d'autres associations courantes. Ainsi on emploie souvent le nom de la capitale à la place du nom de pays : *les négociations avec* ***Ankara*** pour *les négociations avec la* ***Turquie*** ou *le choix de tourner le dos à l'****UE*** *est celui d'****Ankara****, pas celui de* ***Bruxelles***. On désigne la personne par la fonction officielle qu'elle assume (*le président turc* à la place de *Recep Tayyip Erdogan*). De même, une ville peut être interprétée comme représentant sa population ou comme un lieu géographique :

> **Varsovie** vote toujours pour les députés sortants tandis que **Cracovie** se laisse tenter par de nouveaux candidats.
> **Varsovie** n'est pas le centre géographique de la Pologne.

Certains exemples montrent que sans le recours à une base extensive de données recueillies, il est pratiquement impossible de doter un dictionnaire d'informations permettant de prévoir les associations faites dans un texte au nom propre. On ne devrait pas oublier qu'une même entité peut être désignée de multiples façons car il n'y a aucune contrainte qui imposerait de façon obligatoire l'emploi de tel ou de tel autre nom. Ainsi l'actuel président turc peut être qualifié d'*homme fort d'Ankara*, on recourt parfois à la métaphore (*Paris – ville-lumière*), ou même on exploite nos connaissances historiques qui n'ont rien à voir avec les

traits définitoires des mots et devraient être consignées dans une base de données encyclopédiques (*quadruple médaillé d'or lors des Jeux olympiques d'été de 1936 à Berlin, Jesse Owens*). Qui plus est, ces associations se retrouvent en dehors du domaine des ENAMEX. En ce qui concerne les dates, elles se confondent assez fréquemment avec des événements, ce qui est typique pour les noms de cette classe ayant la nature des indices temporels et de ce fait naturellement combinables avec les unités lexicales de nature prédicative : ***Le 4 septembre 476***[4] *marque la fin de l'Empire romain d'Occident*.

En dépit de ces difficultés, ce sont les entités nommées ENAMEX renvoyant à des personnes qui sont les plus faciles à identifier. Le cas le plus simple est celui de noms propres isolés dont la caractéristique est de commencer par une majuscule non précédée par un signe de ponctuation marquant la fin d'une phrase[5]. Ils sont identifiés grâce à un critère formel de typographie, annotés et sortis sans difficultés. Les problèmes d'interprétation surviennent si pour la vedette examinée, le dictionnaire des noms propres consulté au début de l'opération contient plus d'une caractéristique (cas d'ambiguïté). Si, face à une requête pointue, on ne trouve dans le dictionnaire que des étiquettes correspondant à des divisions sommaires, l'identification est possible grâce au terme déclencheur qui joue le rôle de sélecteur. Nous en trouvons un dans chaque entité pluriélémentaire ci-dessous : *le juge Morin*, *l'inspecteur Truchot*, *l'entraîneur Aimé Jacquot*, *le président Emmanuel Macron.* Il est évident que cette liste pourrait être allongée avec d'autres déclencheurs : *personne*, *musicien*, *chanteur*, *sportif*, *politicien*, *acteur* tout en restant dans les limites d'entités nommées.

L'examen de ces entités nommées fait apparaître une structure binomiale : terme déclencheur (TD) + nom propre (NPr) simple ou complexe (= patronyme accompagné éventuellement d'un prénom).

Au plan sémantique, elle est réductible à un prédicat jouant en surface le rôle de terme saillant déclencheur (p. ex. *juge*, *inspecteur*, *président*) suivi de son argument exprimé par un nom de personne. Chaque structure de ce type se laisse convertir en une phrase construite sur *être*, le nom propre s'insérant en position d'argument ouverte par ce verbe conformément aux exigences de celui-ci :

le juge Morin : Morin est (un) juge
l'entraîneur Aimé Jacquot : Aimé Jacquot est un entraîneur
l'inspecteur Truchot : Truchot est un inspecteur
le président Emmanuel Macron : Emmanuel Macron est président

[4] Date de l'abdication de Romulus Augustule.

[5] Cf. aussi Daille et Morin (2000), Friburger (2002). Il s'agit là d'un indice plutôt que d'un critère de définition.

Consacrons un peu d'attention aux termes déclencheurs qui trouvent leur place dans un dictionnaire généraliste où ils devraient être décrits de manière à permettre de catégoriser de façon effective les entités nommées du type décrit ci-dessus[6]. Tout dépend au final de la granularité de description des unités lexicales stockées. Si on se bornait à admettre comme étiquettes caractérisantes [+machine], [+organisation], [+personne], on arriverait, certes, à discriminer les entités nommées telles que *un tracteur John Deer*, *la compagnie John Deer*, *le PDG John Deer* mais on resterait impuissant devant le couple *le PDG John Deer* et *le chanteur John Deer* c'est-à-dire là où il s'agirait de sélectionner les artistes en les départageant d'avec les hommes d'affaires. De même, s'arrêter au niveau des [+mammifères ongulés herbivores] ne permettrait pas de sélectionner, en les bien distinguant, les animaux périssodactyles et les artiodactyles. Plus fine est la classification des termes déclencheurs – et les recherches menées actuellement vont dans ce sens, plus satisfaisants seront les résultats de l'opération de reconnaissance d'entités nommées. Disons cependant que, vu les dimensions du lexique qui devrait être traité de la sorte, il ne semble pas envisageable de pouvoir arriver à un degré de granularité qui garantisse l'infaillibilité de la reconnaissance de toutes les entités nommées dans tous les contextes.

Les substantifs déclencheurs sont sémantiquement complexes et se présentent comme un jeu de poupées russes. Le mot de départ est défini avec des termes qui à leur tour sont définis par d'autres, formant ainsi une hiérarchie d'éléments reliés par une relation d'hypéronymie et correspondant à une ontologie implicite. Ainsi pour *juge* le dictionnaire Larousse mentionne 5 niveaux intermédiaires où les mots expriment chacun un trait définitoire avant d'arriver au terme final :

juge → magistrat → fonctionnaire → officier civil → agent public → personne → être humain

juge – magistrat chargé de rendre la justice en appliquant les lois ;

magistrat – tout *fonctionnaire* ou *officier civil* investi d'une autorité juridictionnelle (membre des tribunaux et des cours, etc.), administrative (maire, préfet, etc.) ou politique (ministre, président de la République, etc.) ;

fonctionnaire – agent public qui, nommé dans un emploi permanent, a été titularisé dans un grade de la hiérarchie des administrations de l'État ;

[6] Des tentatives ont été entreprises d'établir des listes de marqueurs pour différents types de relation. Ainsi Jackiewicz (1996) a étudié la question pour la relation partie-tout tirant profit des verbes *être* et *avoir* ainsi que des constructions génitives.

officier civil – agent public qui, nommé dans un emploi permanent, a été titularisé dans un grade de la hiérarchie des administrations de l'État ;

agent public – personne qui exerce une action d'une certaine sorte, qui joue un rôle déterminant dans la production d'un fait humain ou social ; cause, moteur ;

personne – être humain[7].

À chaque étape, une classe d'objets est définie à laquelle renvoie le nom de personne. En effet, il est possible de convertir l'entité nommée en une paraphrase avec le verbe *être* comme suit :

Morin est un fonctionnaire
Morin est un officier civil
Morin est un agent de l'État
Morin est une personne
Morin est un être humain

Cela étant, devant une requête exigeant d'identifier dans un texte les êtres humains, le système de reconnaissance fournira la réponse correcte lorsque les descriptions des entrées du dictionnaire de consultation auront atteint le niveau représenté dans notre exemple par *fonctionnaire*, *officier civil*, *agent de l'État*, *personne* et bien entendu *être humain.* En effet, dire de quelqu'un qu'il est fonctionnaire implique qu'il est question d'un être humain. Si par contre le jeu d'étiquettes caractérisantes s'arrêtait au niveau d'officier civil, il serait impossible de sortir tous les fonctionnaires.

Il est impossible de prévoir de quel type d'entité nommée on aura besoin pour imaginer un jeu d'étiquettes caractérisantes indispensables que l'on retrouverait avec les déclencheurs potentiels susceptibles de fonctionner comme prédicats sémantiques au sein d'une entité nommée. Or il existe des prédicats qui imposent des restrictions sévères sur leurs arguments : *boire*, *dégouliner*, *couler* connotent un argument marqué [+liquide] en position d'objet pour le premier, en position de sujet pour les deux autres, *miauler* exige un sujet de type [+animé félin], *aboyer* est compatible avec un sujet marqué [+chien], *grouiller* impose un sujet [+pluriel]. Citons enfin un cas extrême. On sait que les traits sémantiques de concours avec les schèmes syntaxiques permettent de différencier les sens des unités lexicales. Or l'analyse du verbe polonais *górować* démontre la nécessité de recourir à des traits sémantiques isolés dans l'ensemble du lexique polonais[8]. Pour

[7] Il est évident que pour compléter la liste de termes déclencheurs, pour chacun des termes de niveau intermédiaire il faudrait tenir compte de leurs synonymes.

[8] Le trait [+arme à feu à tir horizontal] apparaît dans *przenosić* qui est le synonyme de *górować.* Mis à part ces deux verbes, on ne voit pas son utilité dans la description du lexique.

le sens ‘retentir plus fortement que’ (*W chórze górowały głosy dziewcząt* ‘Dans la chorale, c’étaient les voix des jeunes filles qui dominaient’), on a besoin du trait [+son, voix] pour le sujet du verbe, au contraire, le sens ‘se trouver au zénith’ (*W południe słońce góruje* ‘Le soleil se trouve au zénith à midi’) est exprimé si le substantif en position sujet a le trait [+corps céleste], alors qu’il doit être [+arme à feu à tir horizontal] pour exprimer le sens ‘passer au-dessus de l’objectif à atteindre’ (*Ten karabin góruje* – ‘Les obus tirés par ce fusil passent au-dessus de l’objectif à atteindre’).

3. Entité nommée dans le texte

À côté des mécanismes simples de repérage des entités nommées que nous venons de signaler, la porte est grand ouverte à des solutions plus compliquées. Ainsi dans :

> **Une Américaine** a été victime d’une attaque de requin-tigre vers l’île Cocos. **Cette touriste** qui effectuait une sortie de plongée sous-marine, n’a pas survécu aux blessures infligées par l’animal, rapporte The Costa Rica Star.

La reprise de l’entité nommée *Américaine* par *cette touriste* est impossible à détecter uniquement par le recours au dictionnaire généraliste parce que la capacité de faire du tourisme n’est pas le trait définitoire du substantif *Américaine.* Il est cependant évident que le prédicat *touriste* admet parmi ses arguments un nom de nationalité (comme *Américaine*), ce qui dans cet exemple suffit pour garantir l’identité référentielle des deux substantifs.

Si dans l’exemple ci-dessus le lien référentiel est suggéré par le démonstratif *cette* accompagnant le substantif *Américaine*, cette piste ne peut pas être utilisée dans : *Depuis 2012, le requin-tigre est de retour aux environs de cette île du Costa Rica. L’animal avait pourtant disparu pendant près de 30 ans* où on découvre l’identité référentielle entre *requin-tigre* et *animal*, deux entités nommées <animal>.

Conclusion

Il semble que nous atteignons les limites de nos possibilités en ce qui concerne le rendement de nos systèmes de reconnaissance et d’extraction des entités nommées tels qu’ils se présentent aujourd’hui. La voie de l’amélioration de

leurs performances passe par l'augmentation de la granularité de description des vedettes des dictionnaires généralistes alimentés en continu, ce qui est une tâche fastidieuse, et par la mise en veille des pages web où on voit apparaître chaque jour des centaines de mots de toute sorte qui, dûment décrits, doivent élargir le dictionnaire généraliste et la base des noms propres. On attend en dernier lieu une solution nouvelle pour le traitement des métaphores qu'on décèle souvent au sein des entités nommées et qui jusqu'à maintenant sont source d'erreurs dans la reconnaissance d'entités nommées.

Références bibliographiques

DAILLE, Béatrice, MORIN, Emmanuel (2000), « Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations », *Traitement Automatique des Langues*, nº 41 (3), pp. 601-621.

EHRMANN, Maud (2008), *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Thèse de doctorat, Université de Paris XIII.

FRIBURGER, Nathalie (2002), *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques*, Thèse de doctorat, Université François Rabelais de Tours.

JACKIEWICZ, Agata (1996), « L'expression lexicale de la relation d'ingrédience », *Faits de langues*, nº 7, pp. 53-62.

KRIPKE, Saul (1982), *La logique des noms propres*, Paris, Éditions de Minuit.

SAVARY, Agata, PISKORSKI, Jakub (2010), « Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish », in *Intelligent Information Systems. New Approaches* (M. A. Kłopotek *et al.* éds), Siedlce, Publishing House of the University of Podlasie, pp. 141-154.

Sitographie

www.larousse.fr.

www.lemonde.fr/idees/article/2017/09/15/maintenir-le-contact-avec-la-turquie-malgre-erdogan_5186090_3232.html#YxrLmIFsOK7tY9bO.99 (dernière consultation : le 16.12.2017).

www.webmaster-hub.com/publications/les-progres-de-la-reconnaissance-des-entites-nommees-dans-les-moteurs-de-recherche/aaa (dernière consultation : le 20.12.2017).