TALN 2010, Montréal, 19–23 juillet 2010

Détection et résolution d'entités nommées dans des dépêches d'agence

Rosa Stern[1,2] & Benoît Sagot[1]
1. Alpage, INRIA Paris–Rocquencourt & Université Paris 7
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
2. Agence France-Presse – Medialab, 2 place de la Bourse, 75002 Paris, France
rosa.stern@afp.com, benoit.sagot@inria.fr

Résumé. Nous présentons NP, un système de reconnaissance d'entités nommées. Comprenant un module de résolution, il permet d'associer à chaque occurrence d'entité le référent qu'elle désigne parmi les entrées d'un référentiel dédié. NP apporte ainsi des informations pertinentes pour l'exploitation de l'extraction d'entités nommées en contexte applicatif. Ce système fait l'objet d'une évaluation grâce au développement d'un corpus annoté manuellement et adapté aux tâches de détection et de résolution.

Abstract. We introduce NP, a system for named entity recognition. It includes a resolution module for linking each entity occurrence to its matching entry in a dedicated reference base. NP thus brings information relevant for using named entity extraction in an applicative context. We have evaluated NP by the means of a manually annotated corpus designed for the tasks of recognition and resolution.

Mots-clés : résolution d'entités nommées, détection d'entités nommées, extraction d'information.

Keywords: named entity resolution, named entity recognition, information extraction.

1 Introduction

Dans le cadre d'applications utilisant l'extraction d'informations, la détection de segments de texte correspondant à des entités nommées (EN), c'est-à-dire des *mentions* de noms de personnes, de lieux ou d'organisations notamment, n'est pas en elle-même suffisante. Seul un système permettant d'associer à ces mentions un référent extra-linguistique permet d'exploiter le résultat de la détection grâce au sens qui lui est ainsi conféré (Blume, 2005). Comme l'expose en effet avec précision (Ehrmann, 2008), le fonctionnement référentiel des EN constitue une de leurs caractéristiques définitoires. Cette tâche dite de résolution des EN (REN) consiste ainsi à associer à chaque mention d'EN l'entrée adéquate dans un référentiel dédié.

Retrouver le référent correspondant à une mention pose le problème de la polysémie des EN. Comme le rappelle également (Ehrmann, 2008), l'homonymie (*la ville d'Orange* et *l'entreprise Orange*) et la métonymie (*la France* en tant qu'entité géographique vs. en tant qu'organisation dans *La France a gagné la coupe du monde de football*) peuvent en être à l'origine. A cette difficulté s'ajoute celle des différentes variantes qui génèrent une synonymie entre plusieurs mentions se rapportant au même référent (variantes graphiques : *Jacques Chirac* et *J. Chirac* ou surnoms : *Ali le chimique* pour *Ali Hassan al-Majid*). Ainsi, la REN ne peut donc pas s'appuyer sur une correspondance biunivoque entre mention et entrée du référentiel.

Nous proposons un système de détection et de résolution des EN pour le français déployé dans un contexte applicatif que nous présentons à la section 2. Nous décrivons ensuite le système lui-même ainsi que les ressources sur lesquelles il repose (section 3). Nous avons constitué manuellement un corpus annoté (section 4) qui nous permet d'évaluer notre système sur les tâches de détection et de REN de façon intégrée. Les résultats de cette évaluation sont détaillés à la section 5.

2 Contexte d'application

La production d'une agence de presse — dans notre cas, l'Agence France Presse (AFP) — consiste en un flux de données, sous forme de texte mais également de documents multimedia (photo, video, infographie) dont le contenu est informatif à plusieurs niveaux : le traitement des événements médiatiques est doublé d'un ensemble de connaissances relatives à ces événements. Ces connaissances sont en premier lieu des EN, dont les référents sont les acteurs principaux des événements rapportés. Les lieux, personnes et organisations mentionnés dans les dépêches constituent donc des métadonnées dont la collecte permet d'améliorer un certain nombre d'applications de traitement de l'information : indexation des dépêches, filtrage de la production suivant des critères propres aux utilisateurs, tâches de documentation aidées par la recherche d'information. Or les EN deviennent des métadonnées pertinentes lorsque leur statut référentiel est réalisé, c'est-à-dire quand il est possible d'accéder à un ensemble d'informations les concernant et non uniquement à leur mention textuelle. Dans ce contexte, un système ne se limitant pas à détecter des mentions d'EN est particulièrement utile : il permet de rendre ces EN signifiantes grâce à l'identification du référent vers lequel elles pointent.

La production de l'AFP étant généraliste, sa couverture en termes de sujets et domaines est très large ; les EN employées correspondent donc à des références aussi nombreuses que variées : sur les 300 000 dépêches en français de 2009, plus de 3 millions de mentions correspondant à environ 150 000 références différentes ont été détectées par le système présenté à la section 3. Le déploiement de ce système de REN doit ainsi tenir compte du grand nombre et de la variété d'EN qui seront potentiellement à détecter puis à résoudre au fil de la production. Cela implique qu'il s'appuie sur des ressources à la fois importantes et non spécialisées, comme l'explique la section 3. Mais ces ressources ne sont pas les seuls ensembles de références utilisés dans la mise en œuvre de la REN. Un *référentiel* est en effet développé parallèlement afin de centraliser et de structurer les connaissances représentatives du métier de l'Agence. Ce référentiel est basé sur une ontologie rendant compte du modèle de représentation de ces connaissances et dont les instances sont des EN. L'instanciation d'une EN dans ce référentiel se fait si l'on reconnaît à l'entité en question un statut pertinent pour la description des connaissances de l'Agence.

L'interaction entre le système de REN et la production de dépêches est donc double : il s'agit d'une part de résoudre les mentions d'EN détectées au fil des textes en regard des ressources référentielles employées par ce système, et d'autre part d'opérer cette résolution en correspondance avec les instances du référentiel existant. Ce second point implique que les résultats de la REN sont potentiellement plus larges que l'ensemble des instances, et qu'il a donc également pour fonction de proposer de nouvelles entrées au référentiel. Cet aspect évolutif de l'analyse des données est particulièrement important dans le contexte d'une agence de presse : sa production reflète à un rythme rapide les fluctuations, nouveautés et obsolescences de l'usage des EN au cours du temps[1].

[1]Cela vaut d'ailleurs pour l'ensemble du lexique et non seulement pour les EN : les dépêches d'agence suivant l'actualité, les néologismes, nouveaux emplois ou nouveaux types lexicaux (sujets Twitter par exemple) y font régulièrement leur apparition.

3 Détection et résolution des entités nommées : le système NP

NP, notre système de détection et de résolution des EN, fait partie de la chaîne SXPipe, une chaîne robuste et modulaire de traitement de surface qui traite du texte tout-venant dans différentes langues (Sagot & Boullier, 2008)[2]. Différents modules permettent de traiter les problèmes d'encodage et de translittération, de reconnaître différents types d'entités (URL, nombres, dates, adresses, séquences en langue étrangère...) et de corriger l'orthographe et traiter les composés et les amalgames.

NP est constitué de deux modules, l'un pour la détection et le typage et l'autre pour la désambiguïsation et la résolution des EN. Ils reposent sur une volumineuse base d'entités.

Ressources utilisées Notre base de données, décrite plus en détails dans (Stern & Sagot, 2010), contient environ 800 000 entités et 1,3 million de variantes dénotationnelles. Elle comprend des lieux, des organisations, des personnes, des entreprises, des produits et des œuvres (titres de livres, de films, etc.). Nous avons construit cette base à partir de différentes sources, principalement `GeoNames` pour les noms de lieu[3] et la Wikipedia française pour les autres types d'entités[4].

Pour les noms de lieux, nous avons tout d'abord filtré la volumineuse base `GeoNames` [5]. Pour chaque entité conservée, nous avons extrait l'identifiant `GeoNames`, le nom normalisé `GeoNames` (éventuellement complété par un indice permettant de distinguer les entités homonymes), un poids (la population indiquée par `GeoNames`), les variantes dénotationnelles associées, et des indications permettant de visualiser la localité dans *Google maps*[6].

Pour les autres types d'entités, nous nous sommes appuyés sur la Wikipedia française, dans la lignée de travaux antérieurs (Balasuriya *et al.*, 2009; Charton & Torres-Moreno, 2009). Nous avons extrait et typé plus de 170 000 entités à partir des articles Wikipedia sur la base de leur *catégorie*. Le nom normalisé de l'entité est le titre de son article, et différentes variantes dénotationnelles sont extraites au moyen des liens de redirection. Pour les noms de personnes, des variantes supplémentaires sont calculées. Nous identifions tout d'abord le prénom dans les différentes variantes de base puis produisons des variantes avec prénoms abrégés et sans prénoms. Enfin, nous associons à chaque entité un poids, défini par le nombre de lignes de son article Wikipedia.

Détection et typage Une grammaire non-contextuelle de 130 règles a été développée pour détecter et typer les entités nommées à partir de la base ainsi construite ; des motifs contextuels (p.ex. *ville/localité/village de* ou *Dr/M./Mme.*...) sont également utilisés. La reconnaissance est faite de façon ambiguë par l'architecture `dag2dag` de SXPipe (Sagot & Boullier, 2008). Des heuristiques de désambiguïsation sont appliquées pour réduire en partie cette ambiguïté. La sortie de ce module est donc un graphe (DAG) dans lequel chaque EN candidate (empan et type) est représentée par une transition différente.

[2] SXPipe est distribué librement sous licence compatible LGPL (`http://gforge.inria.fr/projects/lingwb/`).

[3] `GeoNames` est librement téléchargeable sur `http://www.geonames.org`.

[4] `http://download.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2`

[5] Nous avons conservé toutes les entrées concernant la France, ainsi que les localités et régions des autres pays si leur population est connue de `GeoNames` et supérieure à 200. Nous en avons exclu les variantes dénotationnelles indiquées comme relevant d'une langue autre que le français ou utilisant des caractères inconnus de la langue française.

[6] Latitude, longitude, et une estimation de l'échelle à utiliser en fonction de la population indiquée par `GeoNames`.

Désambiguïsation et résolution L'objectif de ce module est de résoudre les ambiguïtés d'empan et de type (désambiguïsation) et d'assigner à chaque EN conservée une entrée dans la base (résolution). La combinaison d'un empan, d'un type et d'une référence est appelée une *lecture*. Les tâches de désambiguïsation et de résolution sont effectuées conjointement car elles sont vues toutes deux comme le choix d'une lecture parmi plusieurs lectures possibles.

Notre module de désambiguïsation et de résolution repose sur des informations qualitatives et quantitatives, mais pas sur des techniques d'apprentissage automatique, contrairement à des travaux tels que ceux de (Pilz & Paaß, 2009). Il traite en séquence chaque phrase, en prenant en compte le découpage du corpus d'entrée en documents. Nous définissons un *niveau de saillance* pour chaque entité qui est réévalué à chaque nouvelle phrase d'un même document. Ainsi, la mention d'une entité dans le document augmente son niveau de saillance d'une quantité qui dépend de sa forme de surface[7] et de son poids dans la base[8]. À l'inverse, le passage d'un document à l'autre au sein d'un même corpus divise par deux la saillance de toutes les entités. Ne sont alors conservées que les lectures de saillance maximale, à condition qu'elles soient suffisantes, de façon à éliminer les ambiguïtés concernant les EN dans la phrase[9].

4 Développement d'un corpus d'évaluation

Afin d'évaluer notre système de détection et de résolution d'EN, nous avons sélectionné puis annoté manuellement un ensemble de dépêches de l'AFP. Cette annotation a la forme de marqueurs XML balisant le texte et indiquant pour chaque EN les informations concernant sa mention d'une part (frontières du segment de texte), et son aspect référentiel d'autre part (identifiant unique correspondant au référentiel et associé au type de l'EN). Ce corpus permet donc d'évaluer les performances de détection d'EN mais également celles de la résolution, de façon comparable à (Möller *et al.*, 2004).

Ce corpus est constitué de 100 dépêches contenant en moyenne 300 mots. La table 2 présente la distribution des EN selon leur type et le caractère connu ou non de leur référent au vu des ressources de l'annotateur. Il est disponible librement dans le cadre de la distribution de SXPipe.

Les types d'EN annotés dans ce corpus correspondent à ceux habituellement choisis dans la plupart des tâches organisées par les conférences concernées. Ils sont aujourd'hui limités à *Personnes*, *Lieux* et *Organisations*. L'identifiant reprend celui qui existe dans les ressources de l'annotateur (section 3). Si aucune correspondance de référence ne peut être établie entre la mention et ces ressources, la forme normalisée tient lieu de référence et l'étiquette « inconnue » lui est attribuée. Par ailleurs les annotations de mentions ne comprennent pas les tokens non constitutifs du nom de l'entité lui-même : les titres précédant les noms de personnes notamment en sont exclus (*Mme* ou *Dr*).

```
- Le président <Person name="Barack Obama">Barack Obama</Person> a approuvé un accord
- grippe porcine au <Location name="Canada (2)">Canada</Location> a été révisé
- M. <Person name="Jim Buckmaster" ref="unknown">Buckmaster</Person>, se plaint mercredi
```

TAB. 1 – Exemple d'annotation

[7]Par exemple, une lecture impliquant un nom de personne pour un empan de longueur 1 (nom de famille « nu ») sera défavorisé, surtout s'il s'agirait de la première mention de cette personne dans le document.

[8]La saillance d'un lieu est encore augmentée si elle apparaît dans un document où l'on a détecté une mention de son pays.

[9]Dans le détail, le fonctionnement du module est un peu plus complexe. Il peut notamment préserver certaines ambiguïtés, si une validation manuelle est effectuée en aval (par exemple, pour le développement d'un corpus annoté).

	Lieux			Personnes			Organisations		
	Total	Connus	Inconnus	Total	Connus	Inconnus	Total	Connus	Inconnus
Références	262	218	44	223	111	112	182	99	83
Mentions	673	614	59	424	252	172	447	312	135

TAB. 2 – Distribution des EN sur le corpus selon le type et la référence

5 Evaluation et résultats

Développement et test Le corpus d'évaluation a été divisé en deux parties de tailles équivalentes, correspondant aux données de développement et de test. Une connaissance de l'ensemble des données lors du développement entraînerait en effet une amélioration artificielle des performances du système. Une dépêche sur deux correspond à la partie de test, ce qui permet d'éviter des biais systématiques qui pourraient être causés par certaines EN se trouvant plus fréquemment dans une partie ou une autre du corpus.

Métriques d'évaluation L'évaluation se fait à trois niveaux. D'abord la performance du système est notée en comparant les EN détectées automatiquement en termes d'empan et de type (*Détection et typage* dans la table 3). C'est ainsi que sont construits les scores de la tâche partagée de CONLL 2003 (Tjong Kim Sang & Meulder, 2003). Nous n'établissons donc pas, contrairement à MUC (Grishman & Sundheim, 1996), de scores ou crédits partiels en cas d'erreurs sur l'empan ou sur le type.

Ensuite, pour les mentions dont l'empan et le type sont corrects, nous évaluons la capacité du système à détecter si l'entité correspondante est dans le référentiel ou si elle est « inconnue » (*Détection de la référence*). Pour les mentions qui correspondent à une entité correctement identifiée comme connue, on évalue alors la capacité du système à identifier la bonne entité (*Résolution de la référence*).

Résultats La table 3 présente les résultats obtenus sur l'ensemble de test. La phase de détection a le score le plus bas. Les F-scores pour les types *Personnes*, *Lieux* et *Organisations* sont respectivement 0,85, 0,87 et 0,52 (F-score général de 0,77). Ces chiffres sont satisfaisants lorsqu'on les compare à d'autres systèmes. Évalué sur le même corpus, le système de détection d'entités nommées de Temis, utilisé parallèlement à l'AFP pour l'annotation de dépêches, obtient respectivement 0,86, 0,86 et 0,42. Par ailleurs, les résultats de (Brun *et al.*, 2009) sur leur propre corpus sont respectivement de 0.79, 0.76 et 0.65 (Jacquet, c.p.).

Tâche	Précision				Rappel				F-score			
Type	Lieux	Pers.	Org.	Tot.	Lieux	Pers.	Org.	Tot.	Lieux	Pers.	Org.	Tot.
Détection et typage	0,85	0,92	0,85	0,87	0,88	0,79	0,37	0,70	0,87	0,85	0,52	0,77
Détection de la référence	0,92				0,98				0,95			
Résolution de la référence	0,82				–				–			

TAB. 3 – Résultats d'évaluation pour les 3 sous-tâches : détection et typage des EN, détection de référence, résolution de la référence

Nous avons pu constater, au cours du développement de NP, combien la qualité des résultats est dépendante de celle du référentiel utilisé et des heuristiques mises en œuvre à tous les niveaux. Une amélioration importante des performances de NP est donc envisageable, par exemple en augmentant la couverture du référentiel sur les organisations ou en couplant nos techniques avec des méthodes d'apprentissage.

6 Conclusion et perspectives

Nous avons présenté NP, un système qui complète une phase de détection et de typage des EN par leur résolution, c'est-à-dire le calcul de leur référent. Dans le contexte applicatif du traitement de dépêches AFP, ce système donne des résultats encourageants pour de futurs travaux. Il s'agira notamment d'intégrer des informations directement issues du référentiel ontologique de l'AFP pour aider la désambiguïsation et la résolution, ainsi que des données acquises au cours de la production des dépêches : fréquences d'occurrences des EN, apparition ou disparition au cours du temps. Le développement de ce système pour d'autres langues de travail de l'AFP est envisagé, notamment l'anglais et l'espagnol, avec la constitution de nouvelles ressources référentielles basées sur le modèle construit pour le français.

Références

BALASURIYA D., RINGLAND N., NOTHMAN J., MURPHY T. & CURRAN J. R. (2009). Named entity recognition in wikipedia. In *People's Web '09 : Proceedings of the 2009 Workshop on The People's Web Meets NLP*, p. 10–18, Suntec, Singapour.

BLUME M. (2005). Automatic entity disambiguation : Benefits to ner, relation extraction, link analysis, and inference. *International Conference on Intelligence Analysis*.

BRUN C., DESSAIGNE N., EHRMANN M., GAILLARD B., GUILLEMIN-LANNE S., JACQUET G., KAPLAN A., KUCHARSKI M., MARTINEAU C., MIGEOTTE A., NAKAMURA T. & VOYATZI S. (2009). Une expérience de fusion pour l'annotation d'entités nommées. In *Actes de TALN 2009*, Senlis, France.

CHARTON E. & TORRES-MORENO J.-M. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Actes de TALN 2009*, Senlis, France.

EHRMANN M. (2008). *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. Thèse de doctorat, Université Paris 7 Denis Diderot.

GRISHMAN R. & SUNDHEIM B. (1996). Message understanding conference-6 : a brief history. In *Proceedings of CoLing'96*, p. 466–471, Copenhagen, Denmark.

MÖLLER K., SCHUTZ A. & DECKER S. (2004). Towards an integrated corpus for the evaluation of named entity recognition and object consolidation. In *Proceedings of the SemAnnot Workshop at ISWC2004*, Hiroshima, Japan.

PILZ A. & PAASS G. (2009). Named entity resolution using automatically extracted semantic information. In *Proceedings of LWA 2009*, Darmstadt, Allemagne.

SAGOT B. & BOULLIER P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, **49**(2), 155–188.

STERN R. & SAGOT B. (2010). Resources for named entity recognition and resolution in news wires. In *Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, La Valette, Malte. À paraître.

TJONG KIM SANG E. F. & MEULDER F. D. (2003). Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, p. 142–147, Edmonton, Canada.