Les entités nommées dans le programme QUAERO

Sophie Rosset[1], Cyril Grouin[1], Olivier Galibert[2],
Pierre Zweigenbaum[1], Karën Fort[3,4], Ludovic Quintard[2]
(1) LIMSI–CNRS (2) LNE (3) INIST–CNRS (4) LIPN

1 Introduction

Les entités nommées constituent un champ de recherche actif depuis de nombreuses années, tant du point de vue de leur définition qu'au niveau de la reconnaissance et du typage. Dans le cadre du projet Quaero, nous proposons une nouvelle définition pour des entités nommées étendues en structurant le contenu des entités [1].

Deux corpus de presse ont été annotés sur la base de cette définition. Le premier se compose de données audio transcrites (journaux et émissions de radio et de télévision) tandis que le second intègre des données numérisées à partir d'archives de presse.

Une première campagne d'évaluation a été organisée sur la détection de ces entités nommées étendues sur le premier corpus.

Notre contribution à la journée « *Reconnaissance d'Entités Nommées, Nouvelles Frontières et Nouvelles Approches* » vise à présenter :

- la définition des entités nommées étendues au regard des travaux antérieurs ;
- l'annotation d'un corpus de retranscription de la parole et les mesures de validation de ce corpus ;
- le problème spécifique de la construction d'une référence sur des données transcrites automatiquement ;
- et la campagne d'évaluation liée à ces données.

2 Entités Nommées Étendues

Traditionnellement, les entités nommées sont définies comme relevant des noms propres. Depuis MUC-6 [2, 3], les entités nommées sont classées en trois grandes classes : personnes, lieux et organisations. Des entités de type numériques – dates ou montants – sont souvent considérées comme relevant de la classe des entités nommées.

Des propositions ont été faites pour définir plus finement certaines classes : une catégorie *homme politique* a ainsi été ajoutée à la classe *personne* [4]. Des extensions de la couverture même des entités a également été proposée, telles que les classes *objet manufacturé* et *produit* [5, 6]. Certains travaux ont notamment proposé une extension conséquente des différentes classes avec une hiérarchie de plus de 200 types d'entités [7].

Dans le cadre de nos travaux, en plus d'une extension de la définition des entités nommées (notamment par l'intégration de certains syntagmes nominaux), nous avons proposé une structuration des entités elles-mêmes.

- Ces entités sont hiérarchiques. Elles sont organisées autour de sept types auxquels sont associés entre deux et neuf sous-types. Par exemple, le type <loc> (lieu) a un sous-type <loc.adm> pour les lieux administratifs.
- Elles sont également compositionnelles et comprennent un ou plusieurs composants permettant de typer les éléments constitutifs de l'entité. Certains composants pouvent se retrouver dans différentes entités (ils sont dits transverses, <name> par exemple), d'autres étant spécifiques à un type (par exemple <name.first> pour les entités *personne*).

Suivant cette définition, une annotation manuelle de données orales (retranscription de journaux – parole préparée – et débats – parole spontanée) a été effectuée. Plus d'un million de mots a été annoté. Les taux d'accords inter-annotateurs se situent autour de 0,8 (mesure kappa).

3 Projection de la référence sur la transcription automatique

L'évaluation de la détection de ce type d'entités complexes sur des sorties de systèmes de reconnaissance automatique de la parole (RAP) soulève des difficultés particulières. En effet, puisque leur taux d'erreur est non nul, les mots produits dans leurs sorties peuvent ne pas correspondre à ceux de la transcription manuelle. Comme l'annotation de référence est effectuée sur cette transcription manuelle, il faut trouver une méthode qui projette ces annotations de référence sur les sorties de RAP afin de permettre, d'une part une évaluation fine et fiable, et d'autre part d'offrir à chacun un moyen de visualiser les problèmes posés par ce type de données. Nous avons proposé et implémenté un algorithme permettant cette projection. La difficulté soulevée par cette projection est que la sortie de RAP fournit une annotation de référence des entités nommées correspondant à ce qui a été dit, et non à ce qui est reconnu, ce qui implique une gestion des frontières floues (figure 1).

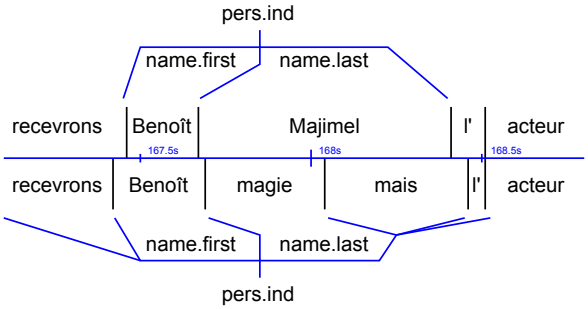

FIG. 1 – Exemple de référence (en haut) projetée sur une sortie de système RAP (en bas)

4 Campagne d'évaluation

Une campagne d'évaluation a été organisée[1]. Sur les transcriptions manuelles, la mesure d'évaluation retenue, le Slot Error Rate, a un maximum de 33%, alors qu'il atteint 61% sur les transcriptions automatiques. Les résultats seront présentés et une discussion concernant l'impact du type de parole (préparée *vs* spontanée) sera proposée.

Références

[1] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, and L. Quintard. Proposal for an Extension of Traditional Named Entities : From Guidelines to Evaluation, an Overview, in *Proc. of LAW*, 2011.

[2] R. Grishman and B. Sundheim. Message Understanding Conference - 6 : A Brief History, in *Proc. of Coling*, 1996.

[3] SAIC. Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.

[4] M. Fleischman and E. Hovy. Fine grained classification of named entities, in *Proc. of COLING*, 2002.

[5] E. Bick. A Named Entity Recognizer for Danish, in *Proc. of LREC*, 2004.

[6] S. Galliano, G. Gravier, and L. Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, in *Proc. of Interspeech*, 2009.

[7] S. Sekine. Definition, dictionaries and tagger of Extended Named Entity hierarchy, in *Proc. of LREC*, 2004.

[1]Ce travail a été organisé dans le cadre du projet Quaero, `http://www.quaero.org/` (financement Oseo, agence française pour l'innovation et la recherche).