Quel est l'apport de la détection d'entités nommées pour l'extraction d'information en domaine restreint ?

Camille Dutrey[1, 2, 3], Chloé Clavel[1], Sophie Rosset[2], Ioana Vasilescu[2], Martine Adda-Decker[2, 3]

(1) EDF R&D, 1 avenue du Général de Gaulle 92141 Clamart

(2) LIMSI-CNRS, rue John von Neumann 91403 Orsay

(3) LPP, 19 rue des Bernardins 75005 Paris

camille.dutrey@{edf,limsi}.fr, chloe.clavel@edf.fr, sophie.rosset@limsi.fr, vasilescu@limsi.fr, madda@{univ-paris3,limsi}.fr

RÉSUMÉ

Les travaux liés à la définition et à la reconnaissance des entités nommées sont généralement envisagés en domaine ouvert, à travers la conception de catégories génériques (noms de personnes, de lieux, etc.) et leur application à des données textuelles issues de la presse (orale comme écrite). Par ailleurs, la fouille des données issues de centres d'appel est stratégique pour une entreprise comme EDF, compte tenu du rôle crucial joué par l'opinion pour les applications marketing, ce qui passe par la définition d'entités d'intérêt propres au domaine. Nous comparons les deux types de modèles d'entités – génériques et spécifiques à un domaine précis – afin d'observer leurs points de recouvrement, via l'annotation manuelle d'un corpus de conversations en centres d'appel. Nous souhaitons ainsi étudier l'apport d'une détection en entités nommées génériques pour l'extraction d'information métier en domaine restreint.

ABSTRACT

What is the contribution of named entities detection for information extraction in restricted domain ?

In the framework of general domain dialog corpora a particular focus is dedicated to Named Entities definition and recognition, which are mostly very generic (personal names, locations, etc.). Moreover, call-centre data mining is strategic for a company like EDF, the public opinion analysis playing a significant role in EDF services quality evaluation and for marketing applications. In this purpose a domain dependant definition of entities of interest is essential. In this primary work we compare two types of entities models (generic and specific to the domain) in order to observe their respective coverage. We annotated manually a sub-corpus extracted from a large corpus of oral dialogs recorded in an EDF call-centre. The respective proportion of generic vs domain-specific Named Entities is then estimated. Impact for future work on building EDF domain-specific entities models is discussed.

MOTS-CLÉS : entités nommées, concepts métier, extraction d'information, données conversationnelles, annotation.

KEYWORDS: named entities, business concept, information extraction, conversational data, annotation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 359–366,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La fouille des données issues de centres d'appel est stratégique pour EDF, compte tenu du rôle crucial joué par l'opinion pour les applications marketing ; elle participe d'une amélioration de la connaissance du client et de ce fait du développement de sa fidélité vis à vis de l'entreprise. Il est de plus nécessaire de mettre en place des méthodes adaptées pour un traitement automatisé afin d'extraire et d'organiser efficacement le contenu informationnel de ces interactions téléphoniques.

L'extraction d'information sur des transcriptions orales est généralement abordée du point de vue de la détection d'entités nommées (EN) et de la recherche d'information, avec des campagnes d'évaluation comme ESTER2 (ESTER2, 2008; Galliano *et al.*, 2009). Ces campagnes sont toutefois principalement basées sur des données de type bulletins d'information et abordent peu la question de l'extraction d'information d'une part sur des conversations téléphoniques, au sein desquelles les spécificités de la parole spontanée sont plus fréquentes, et d'autre part sur des données en domaine restreint. On peut tout de même citer le projet Luna [1], axé sur la compréhension temps-réel de la parole spontanée, et notamment sur la détection de thèmes dans un contexte de dialogue homme-machine (corpus MEDIA, (Bonneau-Maynard *et al.*, 2006)).

Notre objectif est d'explorer la possible utilisation sur des données en domaine restreint de typologies d'EN élaborées pour des données en domaine ouvert, en se basant sur une comparaison entre entités génériques et entités spécialisées. Nous nous focalisons dans cette étude sur des transcriptions manuelles de conversations en centres d'appel. Nous avons pour cela annoté manuellement un corpus en entités génériques structurées définies dans (Grouin *et al.*, 2011) pour le projet Quaero ainsi qu'en entités métier définies par EDF. [2]

Après un aperçu des travaux sur la définition des EN ainsi qu'une description des modèles retenus pour cette étude (section 2), nous décrivons les expériences d'annotation que nous avons menées autour de la comparaison d'entités (section 3). Nous présentons ensuite les résultats de cette étude (section 4) puis nos perspectives de recherche (section 5).

2 Entités nommées génériques et entités d'intérêt spécifiques

2.1 Positionnement

Les EN sont traditionnellement définies en référence aux trois classes développées pour la tâche de reconnaissance d'entités de la conférence MUC-6 [3] : *ENAMEX* (noms de personnes, de lieux et d'organisations), *TIMEX* (expressions temporelles) et *NUMEX* (expressions numériques de pourcentages et de monnaies). Cette première définition a depuis été élargie : ainsi, les EN sont présentées comme suit dans (ESTER2, 2008) : « les EN sont des types particuliers d'unités lexicales (groupes de mots) qui font référence à une entité du monde concret dans certains domaines spécifiques [...] et qui ont un nom (typiquement un nom propre ou un acronyme) ». Cette définition a servi de base à l'élaboration des entités structurées du projet Quaero (Grouin *et al.*,

1. `http://www.ist-luna.eu/project_description.htm`
2. (Danesi et Clavel, 2010) ont abordé la problématique de la détection d'entités métier EDF à partir de transcriptions automatiques de conversations en centres d'appel, ce travail s'appuyant sur les transcriptions manuelles du même corpus. La question de l'impact des erreurs de reconnaissance automatique de la parole n'est pas abordée dans notre étude.
3. `http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html`

2011; Rosset *et al.*, 2011), qui propose d'étendre leur portée à des expressions ne comportant pas de noms propres. (Sekine *et al.*, 2002) proposent une typologie hiérarchique et étendue avec près de 200 catégories. Les trois classes décrites supra ont été enrichies d'autres types d'entités, comme les *produits* (Sekine et Nobata, 2004) ou les *fonctions* (Galliano *et al.*, 2009). La notion recouvrant une réalité sémantico-lexicale de plus en plus large, l'appellation même d'« entité nommée » paraît parfois restrictive. La campagne ESTER2, du fait de l'intégration des *temps* et *montants*, parle d'« entités spécifiques ». Dans cette étude, nous préfèrerons le terme « entités d'intérêt » à « entités nommées ».

(Ehrmann, 2008) analyse la problématique des EN du point de vue théorique des difficultés définitoires et catégorielles, proposant la définition suivante : « étant donnés un modèle applicatif et un corpus, on appelle EN toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Cette définition dépendante d'un corpus fait écho aux travaux de (Boufaden, 2004), celle-ci étudiant l'adéquation d'EN traditionnelles à un corpus spécialisé dans un cadre d'extraction d'information sur des transcriptions de conversations téléphoniques. Cette étude met en avant la différence d'actualisation lexicale entre les EN classiques et les entités d'intérêt en domaine restreint. Nous nous sommes penchés sur l'étude d'une telle approche contrastive des EN dites génériques – dans la mesure où leur définition tend vers une extension sémantique indépendante d'un corpus spécifique et d'un domaine précis – et des entités propres à un domaine en confrontant le modèle Quaero à un modèle issu du domaine du corpus d'étude, la relation entre EDF et ses clients.

2.2 Entités génériques : le modèle Quaero

Le modèle d'entités du projet Quaero, défini dans (Grouin *et al.*, 2011), se différencie des définitions classiques par deux aspects : *i)* les entités sont étendues à des expressions ne contenant pas de noms propres et *ii)* les entités sont structurées grâce à des sous-types et composants [4]. Les différents types et sous-types du modèle, décrits dans (Rosset *et al.*, 2011), sont présentés dans la table 1. Les composants décrits dans le modèle n'ont pas été utilisés pour notre étude.

Types	**Sous-types et/ou Exemples**
Personne	individuelle, collective (*ex. Bertrand Delanoë*)
Fonction	individuelle, collective (*ex. le maire de Paris*)
Organisation	entreprise, administration (*ex. le parti socialiste*)
Localisation	administrative, physique, voie, bâtiment, adresse (*ex. Paris*)
Production humaine	marque, œuvre, production médiatique, produit financier, logiciel, prix, ligne de transport, doctrine, loi/accord (*ex. socialisme*)
Quantité	(*ex. un car de CRS*)
Point dans le temps	date absolue/relative, heure absolue/relative (*ex. le second Empire*)
Événement	(*ex. la coupe du monde*)

TABLE 1 – Entités nommées structurées du projet Quaero

4. Par exemple, l'entité « Bertrand Delanoë » sera typée *personne individuelle* et chacun de ses composants « Bertrand » et « Delanoë » sera respectivement sous-typé *prénom* et *nom*.

2.3 Entités d'intérêt métier : le modèle EDF

Les entités définies à l'EDF correspondent à des concepts métier identifiés par des experts du domaine. Elles ne relèvent pas d'une terminologie globale à l'échelle de l'entreprise, comme le serait une ontologie, mais constituent une multitude de modèles élaborés en fonction *i)* d'un besoin applicatif (formation des conseillers clientèle, analyse marketing...) et *ii)* d'un type de données (champs de commentaires rédigés par le conseiller suite à un appel, enquêtes de satisfaction...). De plus, tout comme les entités Quaero élargissent le champ des EN traditionnelles, les entités d'intérêt EDF recouvrent une réalité lexicale de taille et de composition variables au travers de segments correspondant aux expressions caractérisant les concepts métier. Nous avons sélectionné un modèle d'entités d'intérêt très spécifique élaboré afin d'identifier les motifs des réclamations, des problèmes techniques aux problèmes de contact (présenté en table 2).

Concepts	**Sous-Concepts et/ou Exemples**
Facturation	Index (*ex. facture estimée*) ; Relèvement/Redressement (*ex. facture rectificative*) ; Recouvrement/Relance (*ex. relance pour impayés*) ; Paiement (*ex. mensualisation*)
Incident technique	Assurance (*ex. direct assurance*) ; Panne (*ex. l'ordinateur a grillé*) ; Sinistre (*ex. tempête*) ; Coupure/Surtension (*ex. coupure de courant*)
Intervention technique	Relève (*ex. l'agent de relève*) ; Autre hors relève (*ex. dépannage*) ; Frais d'intervention (*ex. frais de dédit*)
Contact	Attente (*ex. veuillez patienter*) ; Accueil
Traitement de la demande	Délai ; Échanges client-agent
Prix	(*ex. moins cher*)
Contrat/Tarif	(*ex. Jours tempo*)
Vente directe	Offres EDF (*ex. Suivi conso*)
Agence en ligne	(*ex. site Internet EDF bleu ciel*)
Concurrence	(*ex. Powéo*)
Autre	Entités d'intérêt EDF ne relevant pas de la thématique réclamation

TABLE 2 – Entités d'intérêt EDF orientées « réclamation »

3 Expériences

Afin de confronter les deux modèles d'entités décrits en section 2 sur un corpus lié au modèle spécifique EDF, nous avons sélectionné un corpus composé de conversations issues de centres d'appel (section 3.1) que nous avons annoté manuellement en entités Quaero et EDF (section 3.2).

3.1 Le corpus CallSurf

Dans le cadre du projet Infom@gic – CallSurf (Garnier-Rizet *et al.*, 2008), EDF a effectué une campagne d'enregistrement au sein d'un de ses centres d'appel : dix agents volontaires ont été enregistrés durant quatre mois lors de leurs appels avec des clients professionnels. Cette campagne a permis de constituer un large corpus de parole spontanée (5 755 appels, 620 heures de conversations) dont les caractéristiques sont détaillées dans (Danesi et Clavel, 2010). Les auteures ont relevé les spécificités propres à l'oral de ce corpus (par exemple les phénomènes liés au travail de mise en mots et les effets disfluents).

Notre corpus de développement, le corpus Cal10, est composé de transcriptions manuelles produites à l'aide de `Transcriber` (Barras *et al.*, 1998). Le corpus Cal10 est composé de 90 appels téléphoniques soit environ 10h de signal en langue française ; ses principales caractéristiques sont présentées en table 3.

Nous en avons extrait – via un tirage aléatoire – un sous-corpus Cal10a de 10 conversations, soit environ 10 000 mots répartis sur 2 250 tours de parole.

Nombre d'appels	90
Nombre de tours de parole	12 825 (moy. 142,5 / appel)
Nombre de locuteurs	moy. 2,3 / appel
Nombre de mots	100 677 (moy. 1 118,6 / appel – 3,6 / sec.)
Durée des tours de parole	08 : 22 : 16 (moy. 00 : 00 : 02)
Durée totale retranscrite	08 : 51 : 20

TABLE 3 – Caractéristiques générales du corpus Cal10

3.2 Annotation manuelle en entités d'intérêt

Deux annotateurs ont respectivement annoté la totalité du corpus Cal10a en entités Quaero et en entités EDF, avec la plateforme d'annotation `Glozz` (Widlöcher et Mathet, 2009) [5]. L'objectif de cette étude consistant en une comparaison de la couverture d'un modèle d'entités génériques et d'un modèle d'entités spécifiques sur un corpus en domaine restreint, nous avons créé deux modèles d'annotation : l'un pour les entités Quaero et l'autre pour les entités d'intérêt EDF. La figure 1 présente un aperçu de l'interface `Glozz`. Le texte annoté comporte l'entité EDF « il y a pas les heures creuses » (catégorie *contrat*) recouvrant l'entité Quaero « heures creuses » (catégorie *production humaine*).

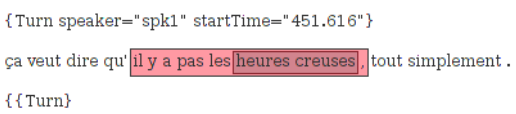

FIGURE 1 – Aperçu de Glozz cadrant un tour de parole annoté

4 Résultats

Suite à l'annotation manuelle du corpus, 368 entités Quaero et 273 entités EDF ont été identifiées dans notre corpus, soit en moyenne 0,3 entités Quaero et 0,2 entités EDF par tour de parole équivalent à une entité Quaero tous les 31 mots environ et une entité EDF tous les 42 mots environ. La figure 2 présente la répartition des entités EDF et Quaero (les entités absentes du corpus ne sont pas apparentes), répartition inégale et ce au sein des deux modèles.

5. `Glozz` permet l'annotation d'unités de granularité variable (caractère, mot, phrase, paragraphe, etc.), de relations entre ces unités et de schémas, constructions complexes de relations.

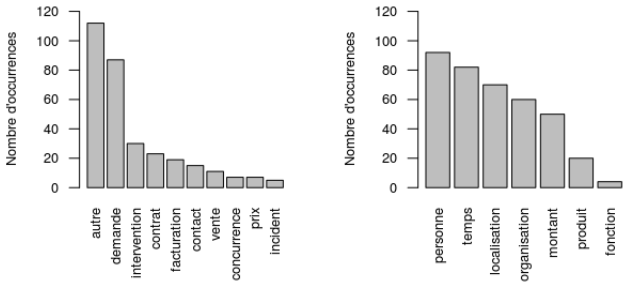

FIGURE 2 – Répartitions des entités EDF et Quaero par catégorie

Concernant les entités d'intérêt EDF, la catégorie *autre* est la plus présente (41% des entités sont de ce type) : cette catégorie concerne des entités d'intérêt pour EDF ne relèvant pas du modèle choisi ; l'on observe ainsi que la thématique de la réclamation clients ne représente qu'une partie des termes métier présents dans le corpus. Au sein du modèle, 32% des entités concernent un *traitement de la demande* ; les autres types d'entités sont relativement peu présents dans notre corpus. Concernant les entités Quaero, la répartition est davantage homogène : environ 25% des entités sont de type *personne*, ce qui est logique compte tenu de la teneur informationnelle du corpus (l'agent est tenu de se présenter ainsi que l'entreprise, et les références client – y compris son nom et le nom de son entreprise – sont fréquemment rappelées) [6]. La forte présence d'entités de type *localisation* (19%) participe également de cette teneur informationnelle. La table 4 présente des informations concernant le chevauchement des deux modèles : le nombre d'entités relevant des deux modèles est relativement faible (respectivement 23,8% des entités EDF et 17,7% des entités Quaero). Ce résultat relève en fait d'un double aspect : certaines entités Quaero sont peu représentées dans notre corpus (*fonction*, *événement*), en raison d'une différence thématique justifiée, alors que d'autres parmi les plus fréquentes (*personne*) ne sont pas prises en compte par le modèle EDF.

Entités EDF ayant au moins un mot en commun avec une entité Quaero	23,8 %
Entités Quaero ayant au moins un mot en commun avec une entité EDF	17,7 %
Entités EDF n'ayant aucun recouvrement avec le modèle Quaero	76,2 %
Entités Quaero n'ayant aucun recouvrement avec le modèle EDF	82,3 %

TABLE 4 – Recouvrement des modèles d'entités Quaero et EDF sur le corpus EDF

Le corpus ayant servi pour l'annotation comporte 11 429 mots, dont 2 413 ont été annotés (soit environ 21% de la totalité du corpus).

Les tables 5 et 6 présentent la couverture des deux modèles d'entités ainsi que leur intersection eu égard à l'ensemble des mots du corpus (table 5) et à l'ensemble des mots annotés dans le corpus (table 6). Bien que davantage d'entités Quaero aient été identifiées, elles concernent seulement 6% des mots annotés ; de plus, elles sont en moyenne composées de 1,8 mot (contre

6. Les données ayant été anonymisées, en accord avec les exigences de la CNIL, l'annotation en noms de personnes a été effectuée grâce à la présence de labels de remplacement tels *NOMPERS*, *NOMSOCIETE*, etc.

6,9 pour les entités EDF). Ces chiffres sont cohérents avec la définition des deux modèles, les entités d'intérêt EDF identifiant des expressions allant au delà du syntagme nominal.

Modèle	Taux de mots annotés
EDF	15 %
Quaero	4,5 %
Intersection	1,5 %
Union	*21 %*

TABLE 5 – Couverture des entités relatives à l'ensemble des mots du corpus

Modèle	Taux de mots annotés
EDF	71,86 %
Quaero	21,59 %
Intersection	6,55 %
Union	*100 %*

TABLE 6 – Couverture des entités relatives à l'ensemble des mots annotés du corpus

Une comparaison des définitions des modèles permet d'établir un rapprochement sémantique entre certaines classes ; nous avons cherché à savoir si ces classes s'actualisaient de la même manière dans notre corpus. En prenant le modèle Quaero comme référence, nous observons notamment que 15,3% des mots annotés *point dans le temps* étaient également annotés *demande/délai* dans le modèle EDF. De même, 23,6% des mots annotés *quantité* relèvent des types EDF *vente* ou *prix*. Enfin, 48,6% des mots annotés *production humaine* relèvent également du type EDF *contrat*. En revanche, seul 1,6% des mots relevant du modèle Quaero (tous types confondus) fait partie de la classe *autre* du modèle EDF.

5 Conclusions & Perspectives

Nous avons mis en avant la présence partielle d'un modèle d'entités génériques comme Quaero dès lors qu'il est appliqué à un corpus de textes issus d'un domaine précis, eu égard aux entités d'intérêt métier pour une entreprise.

Notre étude laisse toutefois apparaître un lien non négligeable entre les deux modèles considérés, grâce au fort recouvrement entre certains types Quaero et EDF : nous souhaitons approfondir cet aspect en étudiant les segments textuels identifiés dans chaque modèle, par types et sous-types. Ces équivalences sont encourageantes pour la suite de nos travaux : nous souhaitons en effet nous pencher sur la détection d'EN en soutient à l'extraction d'entités métier.

Nous continuons cette démarche comparative doublement enrichissante pour affiner les thématiques propres au domaine mais aussi pour élaborer des modèles facilement adaptables aux différents volets du corpus analysé ou à d'autres corpus conversationnels, en mesurant plus finement l'écart observé entre modèle générique et spécifique.

Par ailleurs, un volet essentiel que nous souhaitons étudier est l'impact des erreurs de reconnaissance automatique de la parole sur la détection des entités d'intérêt EDF, dans la continuité des travaux initiés par (Danesi et Clavel, 2010) : un travail de mise en rapport des résultats de la détection d'entités sur les transcriptions manuelles et automatiques représente un objectif crucial du travail futur. Par exemple, de nombreuses erreurs d'analyse concernent les noms propres (personnes, lieux...), comme en témoigne l'exemple suivant : « j'habite à *Beauvoisin* » analysé « j'habite à *vos voisins* ». Nous souhaitons étudier les méthodes existantes de reconnaissance des EN afin de pallier ce type d'erreurs.

Références

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech. *In Proceedings of LREC'98*.

BONNEAU-MAYNARD, H., AYACHE, C., BECHET, F., DENIS, A., KUHN, A., LEFEVRE, F., MOSTEFA, D., QUIGNARD, M., ROSSET, S., SERVAN, C. et VILLANEAU, J. (2006). Results of the French Evalda-Media evaluation campaign for literal understanding. *In Proceedings of LREC'06*.

BOUFADEN, N. (2004). *Extraction d'information à partir de transcriptions de conversations téléphoniques spécialisées*. Thèse de doctorat, Université de Montréal.

DANESI, C. et CLAVEL, C. (2010). Impact of Spontaneous Speech Features on Business Concept Detection : a Study of Call-Centre Data. *In Proceedings of the ACM Multimedia SSCS Workshop*.

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.

ESTER2 (2008). *ESTER2, Convention d'annotation en Entités Nommées, Dates, heures et montants*.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. *In Proceedings of Interspeech'09*.

GARNIER-RIZET, M., ADDA, G., CAILLIAU, F., GAUVAIN, J.-L., GUILLEMIN-LANNE, S. et LAMEL, L. (2008). Callsurf : Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. *In Proceedings of LREC'08*.

GROUIN, C., ROSSET, S., ZWEIGENBAUM, P., FORT, K. et QUINTARD, L. (2011). Proposal for an Extension of Traditional Named Entities : From Guidelines to Evaluation, an Overview. *In Proceedings of Linguistic Annotation Workshop*.

ROSSET, S., GROUIN, C. et ZWEIGENBAUM, P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-CNRS.

SEKINE, S. et NOBATA, C. (2004). Definition, dictionaries and tagger for Extended Named Entity Hierarchy. *In Proceedings of LREC'04*.

SEKINE, S., SUDO, K. et NOBATA, C. (2002). Extended Named Entity Hierarchy. *In Proceedings of LREC'02*.

WIDLÖCHER, A. et MATHET, Y. (2009). La plateforme Glozz : environnement d'annotation et d'exploration de corpus. *In Actes de TALN'09*.