Annotation d'entités nommées sur des transcriptions de parole, le système MAEN

Thierry Poibeau

Laboratoire d'Informatique de Paris-Nord (CNRS UMR 7030 et Université Paris 13)
99, av. J.-B. Clément – 93430 Villetaneuse
thierry.poibeay@lipn.univ-paris13.fr

Résumé

Cet article présente le système d'analyse des entités nommées présenté par le LIPN lors de la campagne Technolangue Evalda Ester 2005. Le système présenté par le LIPN a été initialement développé pour l'écrit [1, 2] et n'a pas subi de transformation de fond pour la campagne Ester. Les principaux ajouts ont permis de compléter les lexiques et la grammaire pour couvrir les catégories proposées dans le cadre d'Ester qui n'était pas couvertes par le système initial. Le système a été appliqué sur les transcriptions de référence ; des tests sur des transcriptions automatiques restent à mener.

1. Introduction

Comme il est dit dans le résumé ci-dessus, le système MAEN (Module d'Analyse des Entités Nommées) a été développé au LIPN pour l'analyse de textes écrits. Ce système repose sur une analyse très classique par transducteurs à nombre fini d'états. La boîte à outil utilisée est le système Unitex, développé à l'Université de Marne-la-Vallée (www-igm.univ-mlv.fr/~unitex/). Le système lui-même repose sur les techniques décrites dans [1] et [2], à savoir essentiellement une analyse lexicale et grammaticale, puis des techniques d'acquisition dynamiques de contextes pertinents, à partir d'une analyse du contexte immédiat de certains mots clés.

Ce document décrit brièvement les principales caractéristiques du système mis en œuvre dans le cadre de la campagne d'évaluation. En plus des dictionnaires généraux, qui permettent de distinguer mots connus et mots inconnus, le système de reconnaissance d'entités nommées est composé de trois éléments principaux :

- des dictionnaires spécifiques,
- des grammaires,
- des processus d'acquisition.

Ces différents éléments sont décrits dans les pages qui suivent, ainsi que les ajouts faits pour la campagne Ester.

2. Les dictionnaires

Différents dictionnaires correspondant aux types d'entités traités ont été mis au point. Les principaux types utilisés, avec l'étiquette Unitex associée et le nombre d'entrées concernées sont donnés dans le tableau 1.

Type	Étiquette	Exemple
Nom de personne	<N+NPR>	Dupont
Prénom	<N+PR>	Jacques
Nom de société	<N+Soc>	L'Oréal
Nom de géographie, dont	<N+Loc>	Picardie
Nom de ville[1]	<N+Loc+City>	Paris
Nom de pays	<N+Loc+Country>	France

Tableau 1.. Contenu du dictionnaire d'entités

L'analyse repose également sur des listes d'amorces et de mots introducteurs (titres, fonctions des personnes dans les entreprises...). Voici une liste des amorces pour les noms de personnes en français :

```
M, Mr, Mrs, Ms, Dr, Mme, Mm, Monsieur,
Madame, Prof, Dr, Lt, Col, LtCol, Sgt,
Maj, Adj, Gén, Cpt, Cpl, Am, SLt, Sdt
```

On s'est par ailleurs servi de lexiques (*Larousse*...) pour définir un ensemble de titres et de noms de profession pouvant introduire des noms de personnes. Pour l'anglais, la grammaire FASTSPEC, liée au système FASTUS et disponible sur le site du

1. La plupart des villes sont indiquées comme nom de lieu, sans autre précision (<N+Loc> et non <N+Loc+City>).

SRI[2], fournit aussi de nombreux titres et mots introducteurs.

3. La grammaire

La grammaire se présente sous la forme d'un ensemble de transducteurs fortement récursif. Les entités considérées sont les noms de personne, de lieux et d'entreprises, ainsi que d'autres éléments comme les dates, les adresses électroniques... Le graphe de niveau supérieur correspond à la figure 1.

Ce graphe se contente de faire appel à des sous-graphes correspondant aux différents types d'entités. Les entités elles-mêmes se décomposent souvent en amorces (*M.* pour *monsieur*) et en composants internes (*Jean* comme prénom, *Dupont* comme nom). Elles peuvent mettre en jeu des mots inconnus ou des ambiguïtés que l'on ne cherche pas obligatoirement à lever (n'importe quel mot commençant par une majuscule après *M.* sera considéré comme un nom de personne).

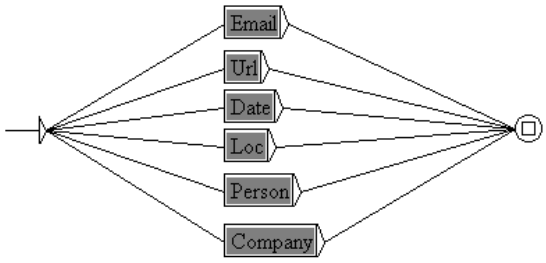


Figure 1 ***Une partie de la grammaire des entités nommées***

La grammaire des entités nommées, une fois compilée, produit un graphe de près de 1000 états et plus de 25 000 transitions. L'algorithme d'analyse fait l'intersection entre le transducteur correspondant à la grammaire et le transducteur du texte préalablement analysé lexicalement. entités). De larges extraits de la grammaire sont donnés en annexe.

4. Les processus d'acquisition

Le système précédemment décrit est enrichi d'un système de généralisation lui permettant d'étiqueter des occurrences isolées de mots inconnus. Le processus utilise les règles dans lesquelles apparaissent des mots inconnus. Par exemple, la séquence *Mr Kassianov* a été reconnue par la grammaire grâce à l'amorce (*Mr*) alors que le mot *Kassianov* était inconnu. Il est possible d'utiliser cette connaissance pour étiqueter dynamiquement les occurrences isolées du mot *Kassianov* comme étant un nom de personne. Cette séquence pourra également être placée dans un dictionnaire intermédiaire qui devra ensuite être validé par l'expert.

Etant donné le type de textes à analyser, cette fonctionnalité a été partiellement débranchée. Seuls certains mots inconnus ont été étiquetés dynamiquement : les mots inconnus entièrement en majuscules ont été étiqueté comme nom de société (type *SNCF*) et les mots inconnus isolés comme mot *unk* (catégorie utilisée en cas de doute sur le type, d'après la documentation Ester).

5. Adaptation du système pour Ester

Le but de l'expérience était pour le LIPN de voir, à titre expérimental, les performances d'un système développé initialement pour l'analyse de documents écrits sur des transcriptions automatiques de la parole. Le système a donc été testé « tel quel », aucune adaptation n'a été prévue pour tenir compte de spécificités de l'oral.

Le corpus d'entraînement a été analysé afin de compléter les données du système initial. Ces données ont dues être filtrées afin d'éliminer les séquences trop ambiguës (mots existants déjà dans les dictionnaires généraux, mots de deux lettres ou moins, *etc.*).

Les autres adaptations apportées au système initial ont concerné l'ajout d'information et de catégories qui n'étaient pas incluses dans le système initial. Ces ajouts ont le plus souvent été symboliques et les catégories non couvertes par le système initial n'ont pas pu obtenir de résultats véritablement significatifs. A titre d'exemple, on peut citer la catégorie *véhicules*. Le système initial ne couvrait pas cette catégorie ; seules les voitures récentes de marque française ont été ajoutées. Il en va de même

2. http://www.ai.sri.com

pour les prix et les récompenses : seuls le prix Nobel et quelques variantes ont été ajoutés.

Les sorties du système ont été modifiées afin de correspondre au format défini par Ester. Là aussi des choix ont été nécessaires afin de faire correspondre les catégories du système initial avec celles de la campagne Ester. Dans la mesure où ces dernières étaient souvent plus détaillées, l'adaptation a été relativement simple.

6. Evaluation

Les résultats officiels et définitifs ne sont pas connus alors que ce document doit être remis (21 mars 2005). De premières tendances laissent à penser d'assez fortes disparités d'un fichier à l'autre (d'une émission transcrites à une autre). Ces variations d'un document à l'autre devront être analysées plus en détail. Il en va de même des résultats d'une catégorie à l'autre : une analyse fine des catégories couvertes par le système initial (dates, noms de personnes et de lieux, *etc*.) devra être menée afin de mettre en évidence les spécificités de l'oral.

Références

[1] Thierry Poibeau. *Extraction automatique d'information, du texte brut au web sémantique.* Hermès-Lavoisier. Paris. 2003.

[2] Jean-François Berroyer. *Tagen, un module d'analyse d'entités nommées*. Mémoire de DEA d'Intelligence Artificielle et Optimisation Combinatoire. Université Paris 13. 2004.

Annexe

Ressources pour la reconnaissance des entités nommées

Cette annexe donne des détails sur les dictionnaires et les grammaires des entités nommées. On détaille ici les catégories faisant traditionnellement partie des entités nommées comme les noms de personnes, de lieux et d'entreprises, auxquels on ajoute les adresses électroniques, les adresses web et les dates. Les ajouts effectués dans le cadre d'Ester ne sont pas mentionnés, qu'ils concernent les dictionnaires ou la grammaire.

Les ressources lexicales

Voici un extrait du dictionnaire de prénoms :

```
Abdallah,.N+PR+Hum:ms
Abel,.N+PR+Hum:ms
Abraham,.N+PR+Hum:ms
Adam,.N+PR+Hum:ms
Adélaïde,.N+PR+Hum:fs
Adolf,.N+PR+Hum:ms
Adolphe,.N+PR+Hum:ms
Adrien,.N+PR+Hum:ms
Agnès,.N+PR+Hum:fs
Ahmed,.N+PR+Hum:ms
```

Un extrait du dictionnaire de noms de personnes :

```
Aalders,.N+NPR
Aara,.N+NPR
Aaren,.N+NPR
Aarika,.N+NPR
Aaron,.N+NPR
Aasen,.N+NPR
Ab,.N+NPRFirst
Aba,.N+NPR
Abadines,.N+NPR
Abagael,.N+NPR
Abagail,.N+NPR
Abana,.N+NPR
Abazari,.N+NPR
Abba,.N+NPR
Abbatantuono,.N+NPR
Abbate,.N+NPR
```

Un extrait du dictionnaire de noms de sociétés (dictionnaire de formes composées) :

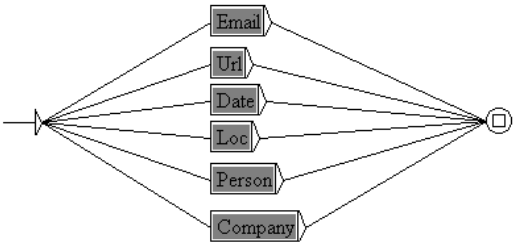
```
A Novo,.N+Soc
AART International,.N+Soc
AB Soft,.N+Soc
Abby Joseph Cohen,.N+Soc
ABN AMRO,.N+Soc
Absolut Vodka,.N+Soc
AC Milan,.N+Soc,.N+Soc
Access Commerce,.N+Soc
Access2net,.N+Soc
Accor Services,.N+Soc
```

Un extrait du dictionnaire de noms de lieux (les noms de ville et de pays sont inclus) :

```
'Idd al Ghanam,.N+Loc
Émirats Arabes Unis,.N+Loc+Country
Öndör Chaan,.N+Loc
États Baltes,.N+Loc
États fédérés de Micronésie,.N+Loc
États-Unis,.N+Loc+Country
Övre Årdal,.N+Loc
Abag Qi,.N+Loc
```

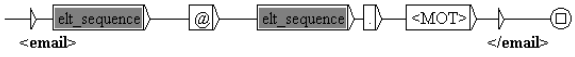
La grammaire des entités nommées

La grammaire des entités nommées est écrite sous la forme d'un transducteur à nombre fini d'états. Ce transducteur est récursif : il s'agit en fait d'un graphe faisant appel à un ensemble de sous-graphes. Chaque sous-graphe traite un type particulier d'entité : noms de personne, de lieu, de société...

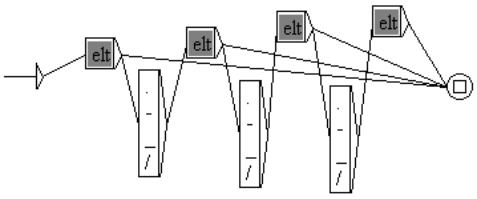


On détaille ci-dessous la grammaire permettant de reconnaître les adresses électroniques, les dates et les noms de personnes. Ces trois catégories dessinent en effet un continuum allant de la reconnaissance de séquences formelles (les adresses électroniques) à des séquences faisant davantage appel à des connaissances sur la langue (les dates ou les noms de personnes).

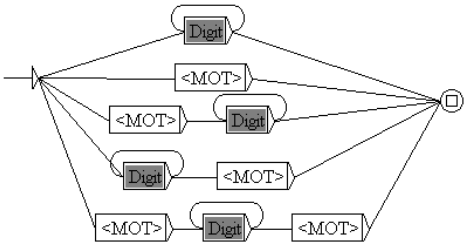
La grammaire permettant de reconnaître les adresses électroniques est simple et ne fait quasiment pas appel à des connaissances de nature linguistique. Les seules étiquettes utilisées sont de type « formel », comme l'étiquette `<MOT>`.



Cette grammaire est fortement récursive. Elle fait appel à un graphe abstrait `elt_sequence`, décrivant l'élément gauche ou droit de l'arobase :

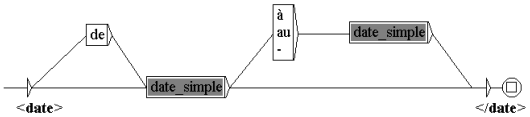


Ce graphe inclut lui-même un autre graphe appelé `Elt` qui est défini comme suit :

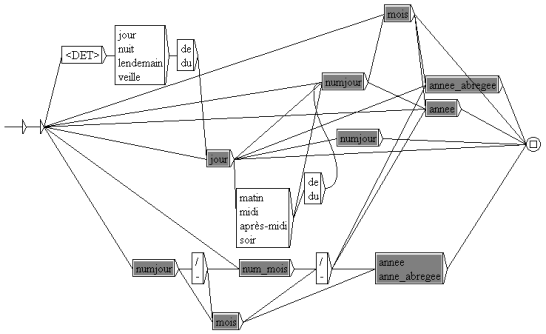


Les deux graphes ci-dessus sont par ailleurs utilisés pour la reconnaissance des adresses web (URL) qui ont une grammaire très proche.

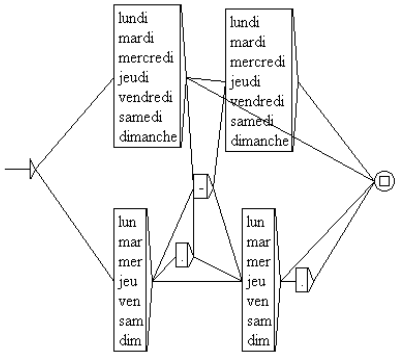
La reconnaissance des dates pose des problèmes particuliers, dus notamment aux intervalles (cf. *de lundi à mardi*, *du 2 au 5 juillet*, *du mardi 2 au jeudi 15*...). La stratégie adoptée consiste à diviser la grammaire en graphes réduits et homogènes, qui permettent de gérer les intervalles à un niveau très local. Ces intervalles sont également traités au niveau syntagmatique pour les cas plus compliqués. La grammaire obtenue est complexe mais reste lisible, malgré le grand nombre de séquences reconnues. Le graphe « supérieur » est le suivant :



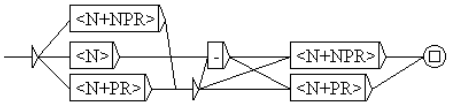
Ce graphe gère un certain nombre d'intervalles. Il fait appel au graphe `Date_simple` :



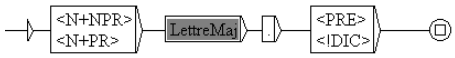
Les sous-graphes permettent de gérer les intervalles et les formes abrégées. Voici par exemple le graphe `Jour` :



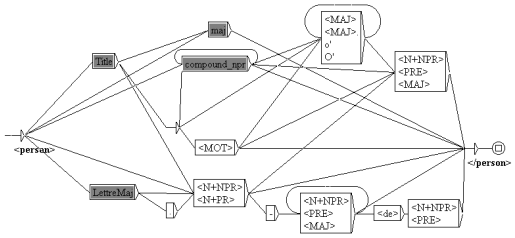
La grammaire des noms de personnes est relativement complexe, du fait de la grande variabilité des séquences possibles. Elle fait appel à des graphes simples comme celui permettant de reconnaître les noms et les prénoms composés (`Compound_npr`) :



ou les noms du type *Thomas J. Jefferson* :



Le graphe global est difficile à lire. Il vaudrait sans doute mieux le réécrire en utilisant davantage de sous-graphes, mais la tâche est plus difficile que pour les dates : il y a moins de « paquets de mots » sémantiquement homogènes pour les noms de personnes que pour les dates (celles-ci se décomposent en effet naturellement en noms de jours, noms de mois...). Voici, à titre indicatif, le graphe supérieur pour les noms de personne :



Application de la grammaire d'entités nommées

La grammaire est appliquée en calculant l'intersection entre les séquences décrites dans les graphes ci-dessus et le graphe du texte. Les sorties des transducteurs (balises `<date>` ou `<person>` figurant dans les graphes décrits) sont insérées autour des séquences reconnues[3]. Ces sorties sont *in fine* transformées pour être mise au format Ester.

[3]. Pour Unitex, cette opération revient à passer un transducteur en mode « fusion » sur le texte.