RECITAL 2005, Dourdan, 6–10 juin 2005

Utilisation de la Linguistique Systémique Fonctionnelle pour la détection des noms de personnes ambigus

Thomas Heitz
LRI - Université Paris Sud
91405 Orsay Cedex - France
heitz@lri.fr

Mots-clefs – Keywords

Linguistique Systémique Fonctionnelle, Détection des entités nommées, Fouille de textes
Systemic Functional Linguistics, Named entity recognition, Text mining

Résumé - Abstract

Dans cet article, nous nous proposons de construire un lexique étiqueté selon les principes de la Linguistique Systémique Fonctionnelle (LSF) et de l'appliquer à la détection des noms de personnes ambigus dans les textes. Nous ne faisons pas d'analyse complète mais testons plutôt si certaines caractéristiques de la LSF peuvent permettre de compléter les modèles linguistiques actuellement utilisés pour la détection des entités nommées. Nous souhaitons ainsi apporter une contribution à l'application du formalisme LSF dans l'analyse automatique de textes après son application déjà éprouvée à la génération de textes.

In this paper, we propose to build a tagged lexicon according to the Systemic Functional Linguistics (SFL) principles and to apply it to the recognition of ambiguous person names in texts. We do not achieve a complete analysis but rather test if some characteristics of the SFL could enable the completion of actual linguistic models used in named entity recognition. We want thus to bring a contribution to the application of the SFL model in automatic text analysis while it already proved its usefulness for text generation.

1 Introduction

Nous allons discuter de l'intérêt du formalisme de la Linguistique Systémique Fonctionnelle (LSF) pour la détection des noms de personnes ambigus dans les textes. Nous nous situons donc dans le domaine de la détection des entités nommées qui est une des tâches essentielles de la fouille de textes (Daille & Morin, 2000).

La LSF (Halliday & Matthiessen, 2004) offre une approche complémentaire des linguistiques habituellement utilisées. La LSF part du contexte social et dès le départ cherche à comprendre la fonction des mots alors que traditionnellement les linguistiques partent des mots pour arriver finalement au sens.

Nous souhaitons ici étudier l'intérêt de caractéristiques de la LSF dans le but de les incorporer dans les dernières étapes d'un système de détection d'entités nommées similaire à ceux utilisés lors de la compétition MUC-7 (Mikheev *et al.*, 1998).

Dans le domaine de la détection des entités nommées, il a été constaté que le problème de la poly-catégorisation d'une entité nommée nécessite une meilleure analyse du contexte pour la traiter (Daille & Morin, 2000). La poly-catégorisation existe, par exemple, pour le nom ambigu de personne *France*, qui peut aussi être un nom de pays.

Nous nous intéresserons ici aux verbes appartenant aux contextes des noms de personnes. Plus précisément, nous chercherons les processus de la LSF associés à ceux-ci puisque les processus sont portés par les verbes. Les processus d'une phrase étant les éléments qui désignent la ou les actions en cours entre les différents acteurs participant à l'action et pour des circonstances données. Puis, nous constaterons si des liens existent entre catégories de processus et contextes de noms de personnes. Ainsi, nous pensons pouvoir améliorer la discrimination entre les noms de personnes et les autres noms propres ou les mots communs dans les cas ambigus.

Nous allons établir une méthodologie pour la détection des noms de personnes ambigus à l'aide des processus de la LSF sur une première partie d'un corpus. Puis, nous validerons expérimentalement notre hypothèse d'amélioration de la détection des noms de personnes ambigus sur une autre partie du même corpus.

1.1 Extraction des noms de personnes

Comme la recherche sur les entités nommées l'a déjà établi (Poibeau, 2001), les noms de personnes sont souvent entourés par des mots inclus dans l'entité nommée. A savoir, les noms de fonction ou de titre qui peuvent se situer en préfixe comme *Mr.*, en infixe comme *de la* ou en suffixe comme *Second du nom*. Par exemple *Mr. Arnaud de la Tour Second du nom*. De nombreuses grammaires et lexiques ont été développés pour repérer ces formes.

En plus de ces mots inclus dans le nom de personne, que nous appellerons indices locaux, les indices contextuels ont été aussi étudiés. Par exemple, la préposition *chez* introduit presque toujours une personne. De même, des grammaires utilisant ces indices ont été développées. Mais ces indices contextuels restent la plupart du temps limités aux mots grammaticaux que sont les déterminants, conjonctions et prépositions pour les règles non apprises. Nous voulons donc l'étendre aux mots lexicaux que sont les verbes, noms, adjectifs et adverbes afin de mieux traiter les cas ambigus de noms de personnes. Dans cet article, nous nous limiterons aux verbes appartenant aux contextes des noms de personnes.

1.2 Processus dans la Linguistique Systémique Fonctionnelle

Les processus appartiennent à la fonction idéationnelle de la LSF (Halliday & Matthiessen, 2004) qui est un modèle d'analyse de la phrase qui divise celle-ci en processus, participants et circonstances. Dans ce modèle, les verbes peuvent être catégorisés dans des processus puisqu'ils les supportent. Il existe six principaux processus : matériel, mental, verbal, relationnel, comportemental et existentiel. Dans le tableau 1 sont caractérisés les participants des différents processus de la LSF.

Processus	Sens	Nombre de participant	Nature du premier participant	Nature du second participant
Matériel	faire et avoir lieu	1 ou 2	chose	chose
Mental	resentir	2	chose consciente	chose ou fait
Verbal	dire	1	source de symboles	
Relationnel	être et avoir	1 ou 2	chose ou fait	
Comportemental	se comporter	1	chose consciente	
Existentiel	exister	1 ou 0[1]	chose ou fait	

Table 1: Caractéristiques des participants des processus de la LSF

Les personnes, en tant que premiers participants, sont impliquées presque systématiquement dans les processus mentaux, comportementaux et verbaux comme chose consciente pour les deux premiers et source de symbole pour le troisième et d'une façon plus incertaine dans les autres processus. Nous allons donc tester l'implication de ces processus dans les contextes de noms de personnes sur un corpus afin de retenir les catégories pertinentes.

2 Méthodologie pour la détection des noms de personnes ambigus à l'aide des processus de la LSF

2.1 Extraction des relations syntaxiques et des noms de personnes

Nous avons utilisé un sous ensemble du corpus Aquaint d'une taille d'environ 10 mégaoctets provenant de la compétition TREC Novelty 2004 (Soboroff, 2004). Il contient des dépêches écrites en anglais en provenance d'agences de presse.

Un lexique de verbes étiquetés selon les processus de la LSF à été élaboré manuellement à partir de (Halliday & Matthiessen, 2004) et étendu en rajoutant l'intersection des hyponymes et synonymes contenus dans WordNet (Miller *et al.*, 1990) pour chaque verbe initial. Puis, nous avons extrait à partir du corpus une liste de noms de personnes de manière semi-automatique avec des listes de noms de personnes déjà connus et une correction manuelle.

Le logiciel Link Grammar 4.1 (Sleator & Temperley, 1993) a été utilisé afin d'extraire les relations syntaxiques des phrases du corpus. Seules les relations syntaxiques de type Sujet-Verbe dénommées S et SI dans la nomenclature de Link Grammar ont été prises en compte.

[1]Par exemple, dans la phrase : *There was a storm.* il n'y a pas de participant juste un processus.

Nous avons ensuite lemmatisé les verbes, étiqueté les verbes lemmatisés selon les processus de la LSF et étiqueté les noms de personnes. Une partie du corpus contenant des noms de personnes ambigus a été mise de côté afin de pouvoir tester les hypothèses faite dans la section 2.2.

2.2 Établissement de correspondances entre les processus de la LSF et les contextes de noms de personnes

Nous avons cherché les types de processus qui qualifient les verbes dont les sujets sont des noms de personnes à partir des relations Sujet-Verbe trouvées précédemment dans la section 2.1. Les résultats sont présentés dans le tableau 2.

Processus	Sens	Lexique	Relation Personne-Verbe	Comparaison
Matériel	faire et avoir lieu	46% (2462)	71% (6627)	+25
Mental	resentir	12% (653)	66% (6098)	**+54**
Verbal	dire	7% (353)	49% (4557)	**+42**
Relationnel	être et avoir	32% (1698)	71% (6609)	+39
Comportemental	se comporter	3% (166)	8% (345)	+5
Existentiel	exister	2% (129)	14% (1305)	+12
Total		5353 verbes	9285 relations	

Table 2: Occurrences des processus dans les relations Personne-Verbe. Il existe une catégorie par ligne dans le lexique mais plusieurs par ligne dans les relations.

Nous pouvons constater que les processus de types mental et verbal semblent les plus adaptés pour trouver les contextes de noms de personnes puisque ce sont les catégories de processus dont la proportion augmente le plus par rapport à celle du lexique utilisé. Ceci confirme en partie ce qui était attendu d'après le modèle théorique exposé dans la section 1.2. Cependant la catégorie relationnel augmente aussi beaucoup bien que les verbes obtenus ne soient pas très pertinents. Il sera donc utile dans une prochaine expérimentation de tenir compte des sous catégories pour obtenir des résultats plus précis.

3 Validation expérimentale

3.1 Ambiguïtés entre noms de personnes et autres noms propres

De nombreux noms de personnes sont aussi des noms d'organisations ou de lieux notamment. Plus généralement, ceci rejoint le problème de la poly-catégorisation des entités nommées.

Dans le tableau 3, sont présentés quelques cas d'ambiguïté avec les autres noms propres. Les résultats sont donnés sous forme de fractions comportant au numérateur le nombre de mots bien identifiés comme un nom de personne ou comme un autre nom propre et au dénominateur le nombre d'occurrences de ce mot pour lesquels les verbes dont il est le sujet ont été catégorisés. Les verbes sont catégorisés selon les processus de la LSF et **lorsqu'ils appartiennent à au moins deux types parmi mental, verbal et comportemental nous considérons que le mot**

est un nom de personne et non une organisation ou un lieu. Les modaux ne sont pas pris en compte.

Type d'ambiguïté	Nom	Précision
Personne - Organisation	Ford	24/35
	Bell	0/5
	Morris	0/0
Personne - Lieu	Virginia	0/0
	Madison	0/0

Table 3: Résultats pour la distinction entre noms de personnes et autres noms propres

3.2 Ambiguïtés entre noms de personnes et mots communs

Il existe de nombreux cas où la distinction entre noms propres et mots communs ne peut s'effectuer grâce à la casse de la première lettre du mot. Les principaux cas sont : premier mot de phrase ou de citation, titre, transcription de l'oral vers l'écrit tout en minuscules, langue sans différenciation de casse entre noms propres et mots communs. En allemand, par exemple, les mots communs s'écrivent avec une majuscule en première lettre comme pour les noms propres. Ceci justifie de traiter les cas d'ambiguïtés entre noms de personnes et mots communs.

Une phrase exemplaire tirée de notre corpus est la suivante : **Ray** *believes the electron beam process is superior to gamma* **ray** *irradiation because the technology is easier and cheaper to implement.*

On peut y voir le mot *ray* sous la forme de nom de personne et de mot commun. La majuscule de début de mot ne suffit pas ici à les distinguer car *Ray* est le premier mot de la phrase qui par définition possède toujours une majuscule en première lettre. Le verbe *to believe*, croire, qui le suit est classé dans les processus de type mental et correspond bien à une catégorie de processus appartenant à des contextes de noms de personnes comme vu dans la section 2.2.

Dans le tableau 4, trois des cas les plus fréquents d'ambiguïtés entre noms de personnes et mots communs sont présentés. Les résultats sont donnés sous la même forme que les résultats précédents et avec le même protocole.

Type d'ambiguïté	Nom	Précision
Personne - Nom commun	bill	49/66
	bird	0/2
	stone	3/3
	ray	4/4
Personne - Verbe	drew	0/0
	mark	0/0
Personne - Adjectif	brown	11/11
	gray	20/32

Table 4: Résultats pour la distinction entre noms de personnes et mots communs

Exemples de phrases traitées :

The **bill died** *in a Transportation subcommittee, but Moreno said it had widespread support in the house [...]* To die, processus mental impliquant un nom de personne, mal catégorisé.

Mel-Daniels was an assistant coach at State when **Bird led** *the team the NCAA final in 1979.* To lead, processus verbal impliquant un nom de personne, bien catégorisé.

On trouve notamment le problème de la poly-catégorisation comme pour le verbe *to die* qui ne devrait pas être classé mental mais matériel et très rarement comme processus mental. Beaucoup de mots communs sont ici personnifiés comme pour le mot *bill, projet de loi* en français, ce qui peut être source d'erreurs.

4 Conclusion et perspectives

Nous nous sommes ici intéressé aux verbes dont le sujet est un nom de personne. D'après ces premiers résultats, même s'il apparaît important de chercher quelle action subit ou effectue un être conscient pour découvrir tous les noms de personnes d'un texte, il semble nécessaire d'avoir des informations supplémentaires comme les autres participants des processus et les circonstances. L'utilisation des sous catégories de processus couplée à l'apprentissage de règles de classification semble aussi une piste à suivre.

Une comparaison avec d'autres méthodes d'expansion de termes pour notre lexique de base et d'autres lexiques catégorisant les verbes différemment serait utile. De même qu'une comparaison avec d'autres systèmes de détection d'entités nommées sur des corpus de référence une fois l'analyse des processus de la LSF intégrée dans un système de détection d'entités nommées.

Je remercie mon directeur de thèse, Yves Kodratoff, pour m'avoir fait découvrir la LSF.

Références

DAILLE B. & MORIN E. (2000). *Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations*, In *Traitement automatique des noms propres sous la direction de Denis Maurel et Franz Guenthner*, chapter 1, p. 601–621. Hermes. Collection : Traitement automatique des langues.

HALLIDAY M. A. K. & MATTHIESSEN C. M. I. M. (2004). *An Introduction to Functional Grammar*. Hodder Arnold. 3nd. edition.

MIKHEEV A., GROVER C. & MOENS M. (1998). Description of the ltg system used for muc-7. In *Message Understanding Conference Proceedings, MUC-7*.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography 3 (4)*, p. 235–244. revised august 1993.

POIBEAU T. (2001). Deconstructing harry, une évaluation des systèmes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information*.

SLEATOR D. & TEMPERLEY D. (1993). Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*.

SOBOROFF I. (2004). Overview of the trec 2004 novelty track. In *NIST Special Publication: The Thirteenth Text Retrieval Conference (TREC 2004)*, p. 57–70.