Extraction de noms propres à partir de textes variés : problématique et enjeux

Leila Kosseim[1] et Thierry Poibeau[2]
[1]RALI, Université de Montréal, kosseim@iro.umontreal.ca
[2] Thales/TRT et LIPN, Institut Galilée, Université Paris-Nord.
Thierry.Poibeau@thalesgroup.com

Résumé – Abstract

Cet article porte sur l'identification de noms propres à partir de textes écrits. Les stratégies à base de règles développées pour des textes de type journalistique se révèlent généralement insuffisantes pour des corpus composés de textes ne répondant pas à des critères rédactionnels stricts. Après une brève revue des travaux effectués sur des corpus de textes de nature journalistique, nous présentons la problématique de l'analyse de textes variés en nous basant sur deux corpus composés de courriers électroniques et de transcriptions manuelles de conversations téléphoniques. Une fois les sources d'erreurs présentées, nous décrivons l'approche utilisée pour adapter un système d'extraction de noms propres développé pour des textes journalistiques à l'analyse de messages électroniques.

This paper discusses the influence of the corpus on the automatic identification of proper names in texts. Rule-based techniques developed for the newswire genre are generally not sufficient to deal with larger corpora containing texts that do not follow strict writing constraints. After a brief review of the research performed on news texts, we present some of the problems involved in the analysis of informal texts by using two different corpora (the first one composed of electronic mails, the second one of hand-transcribed telephone conversations). Once the sources of errors have been presented, we then describe an approach to adapt a proper name extraction system developed for newspaper texts to the analysis of e-mail messages.

Mots-Clefs : Extraction d'information, Entités nommées

1 Introduction

L'identification de noms propres au sein de documents écrits ou oraux est une tâche importante pour le traitement automatique de la langue naturelle. En effet, ce type de syntagme occupe une place prépondérante dans de nombreux corpus. Il est dès lors primordial de pouvoir reconnaître ces expressions tant pour des besoins spécifiques (ex. identifier les différentes fonctions qu'a occupé Lionel Jospin) que plus généraux (ex. améliorer l'analyse syntaxique).

De nombreux travaux ont porté sur l'identification de noms propres dans des textes journalistiques, notamment les *Message Understanding Conferences* (MUC) (MUC95; MUC98). Cette tâche est définie comme étant *générique*, dans la mesure où tous les textes font usage de noms propres et que leur repérage semble a priori reproductible. Cependant, les expériences menées dans le cadre des MUC ont porté sur des corpus homogènes, constitués essentiellement

d'articles de journaux ou de dépêches d'agences. Ce type de textes respecte des contraintes rédactionnelles fortes qui facilitent la tâche de repérage (ainsi, des séquences telles que *M.* pour *monsieur* ou *Mme* pour *madame* précèdent assez systématiquement les noms de personne). Les stratégies utilisées sont toutefois insuffisantes pour analyser d'autres types de textes comme des courriers électroniques ou des minutes de réunion car ces textes ne répondent pas à des critères rédactionnels stricts sur lesquels se fondent la majorité des systèmes à base de règles. Avec l'explosion de textes sur supports électroniques, c'est souvent à des textes issus de flux informels que se trouvent confrontés les systèmes.

2 Travaux antérieurs

Sous l'influence des conférences MUC, les travaux sur l'extraction d'entités nommées (de l'anglais *named entities*) ont traditionnellement portés sur des textes journalistiques. L'identification des entités nommées inclut trois types d'expressions : les noms propres (ENAMEX), les expressions temporelles (TIMEX) et les expressions numériques (NUMEX). Dans ce travail, nous nous sommes concentrés sur le premier type d'expressions : les noms propres.

La reconnaissance des entités nommées à partir de textes écrits est actuellement la tâche d'extraction d'information la mieux réussie. Les taux combinés de précision et de rappel sont comparables à ceux des humains, avec un taux de l'ordre de 0,90 de P&R sur des dépêches journalistiques. Deux grandes approches sont généralement suivies pour leur identification : une approche linguistique "de surface" et une approche probabiliste. L'approche linguistique est fondée sur la description syntaxique et lexicale des syntagmes recherchés. Des règles de grammaire utilisent des marqueurs lexicaux (ex. *Mr* pour *Mister* ou *inc.* pour *incorporated*), des dictionnaires de noms propres et des dictionnaires de la langue générale (essentiellement pour repérer les mots inconnus) pour repérer et typer les syntagmes intéressants (ABD+95; AM99; AHB+95). De son côté, l'approche probabiliste utilise un modèle de langage entraîné sur de larges corpus de textes pré-étiquetés. Cette approche est particulièrement robuste lorsque les textes sont bruités, c'est pourquoi la grande majorité des systèmes dédiés à l'oral adoptent une approche probabiliste (ex. (KSSW98)).

La présente étude vise à étudier la robustesse des systèmes à base de règles sur des corpus non journalistiques[1] car les auteurs ont chacun développé un système à base de règles (Exibum (KL98) et Lexis (Poi99)) ayant la vocation d'être utilisés sur des types de textes variés.

3 Expériences sur des corpus non-journalistiques

Deux corpus d'échanges informels ont été utilisés pour évaluer la robustesse des systèmes : le corpus Valcartier et le corpus Communications. Le corpus Valcartier est composé de transcriptions manuelles de conversations téléphoniques en anglais fournies par la division de Recherche et Sauvetage des Forces Armées canadiennes. Même s'il s'agit de transcriptions de l'oral, la qualité des transcriptions permet de les assimiler à des textes écrits. En effet, les transcriptions sont en casse mixte, contiennent très peu d'erreurs de reconnaissance de mots[2] et contiennent des marques de ponctuation ajoutées par les agents de transcription. Ces caractéristiques nous permettent d'avoir des textes *propres* en ce qui concerne la transcription. Le corpus comporte environ 25 000 signes, soit 2200 mots. Le corpus Communications est composé d'un ensemble

[1] Les approches probabilistes et numériques ne seront donc plus évoquées par la suite.

[2] Contrairement à des textes issus de transcriptions automatiques qui peuvent contenir de 30 à 40% de taux d'erreur de mots.

	MUC-6	Valcartier	Communication
Annotateurs humains	0,97	0,97	0,90
Alembic	0,86	0,50 - 0,57	-
TextPro	0,86	0,41	-
Exibum	0,69	0,44	-
Lexis	0.90	-	0,50

TAB. 1: P&R de l'extraction des noms propres sans adaptation de système

de courriers électroniques en anglais portant sur le domaine des télécommunications. Le corpus est en casse mixte, il est écrit dans un style que l'on peut qualifier d'*informel*, au sens où les phrases peuvent ne pas être complètes mais être rédigées dans un style télégraphique. Finalement, le nombre de fautes d'orthographes est relativement limité au regard d'autres expériences portant sur des corpus de courriers électroniques. Le corpus comporte environ 300 000 signes, soit 50 000 mots.

3.1 Constat initial : Une chute de performances importante

L'utilisation brute de stratégies développées pour des textes journalistiques sur des textes d'une autre nature entraîne une chute considérable des performances. Ce constat a été fait par les deux auteurs dans le cadre d'expériences indépendantes : Valcartier–Exibum et Communication–Lexis. Trois types de noms propres ont été évalués : les noms de personnes, les noms de lieux et les noms d'organisations. Comme l'illustre la table 1, les annotateurs humains ayant participé à ces expériences ont obtenus des taux comparables à ceux des expériences MUC, alors que les systèmes ont obtenus des scores nettement plus faibles. Pour vérifier que ce constat d'échec n'était pas dû à une faiblesse de nos systèmes, nous avons élargi l'évaluation en utilisant Alembic (ABD+95) et TextPro (AM99), deux systèmes disponibles publiquement et qui se classent parmi les plus performants aux MUC. Les résultats de la table 1[3] montrent clairement que tous les systèmes ont des scores bien plus faibles et sont loin de rivaliser avec les annotateurs humains. Il devient donc intéressant d'identifier pourquoi des scores aussi faibles sont obtenus.

3.2 Une grammaire faite de variantes

La syntaxe variable des noms propres est à l'origine de la plus grande partie des cas d'erreurs de silence (i.e. noms propres non détectés). Dans les textes journalistiques, les noms de personnes sont généralement précédés de titre ou d'amorces (ex. *M.*, *Mme*) ; en revanche dans les deux corpus analysés, cette pratique est rare. Comme les noms (surtout les noms de familles) appartiennent à des classes lexicales ouvertes, ces marqueurs sont des indicateurs très efficaces. La grammaire des noms de personnes laisse donc apparaître une géographie originale et mouvante suivant le corpus. Les règles qui s'appliquent sur des corpus issus d'échanges informels ne sont pas obligatoirement différentes de celles d'un corpus journalistique. En revanche, une règle fréquente dans un corpus journalistique pourra être très peu représentée dans un corpus

[3]Notons que le score des annotateurs humains/MUC-6 est le score officiel de l'extraction de toutes les entités nommées (MUC95). Aussi, le score Alembic/MUC-6 est le score officiel (MUC95), alors que le score Alembic/Valcartier a été calculé avec Alembic Workbench 4.12. Dans le cas d'Alembic/Valcartier, deux mesures sont données car deux mots fréquents dans le corpus étaient systématiquement mal étiquetés par le système, le désavantageant injustement. Nous donnons alors une première évaluation avec le corpus original et une seconde évaluation sans tenir compte de ces mots. Finalement, notons qu'Alembic a très mal étiqueté le corpus Communication (au point de ne pas pouvoir être évalué par les outils fournis avec le corpus MUC-6 ; une évaluation informelle donne une F-mesure inférieure à 0,30). Nous avons donc choisi de ne pas faire figurer de score pour Alembic/Communication mais cette expérience confirme nos autres résultats. Pour raisons matérielles, TextPro n'a pas été évalué sur le corpus Communication.

composé de courriers électroniques, et réciproquement pour d'autres règles. La grammaire des noms d'organisation est également plus relâchée dans des corpus informels que dans des corpus journalistiques. Des marqueurs lexicaux (comme *inc.* ou *Ltd.*) sont généralement absents et quelquefois des mots entiers sont omis. Par exemple, le nom de l'organisme *Transportation Safety Board* comprend l'indicateur *Board* dénotant la présence d'un nom d'organisme. Dans le corpus Valcartier, le nom de cet organisme est souvent abrégé en *Transportation Safety* (sans marqueur). Finalement, les noms de lieux sont aussi identifiés par des marqueurs (ex. la préposition *in* ou le mot *city*). L'absence de marqueur crée des erreurs de silence ou de mauvaise catégorisation car le système a alors affaire à des mots inconnus isolés. Seules des stratégies de typage dynamique mettant en jeu l'analyse d'autres occurrences peuvent amener à résoudre ces cas difficiles.

4 Vers des systèmes adaptables

En fonction des observations ci-dessus, un ensemble de stratégies d'adaptation de système ont été mises au point puis évaluées sur le corpus Communication.

4.1 Stratégies d'adaptation

L'absence de marqueur conduit, dans les cas extrêmes, à des mots inconnus isolés. Pour réduire ce problème, il est nécessaire d'augmenter la couverture des dictionnaires et d'adjoindre au système d'extraction des processus d'acquisition dynamique.

Augmenter la couverture du dictionnaire – Quel que soit le corpus sur lequel a porté l'expérience, la grammaire des noms reste relativement stable et la séquence *Prénom Nom* est généralement la plus représentée. Les variantes de cette séquence sont un nom isolé, ou précédé d'initiales. Ce type de séquences précédé d'un marqueur est en revanche quasi absent dans le corpus Communication. Il est donc essentiel d'avoir une bonne couverture des classes semi-fermées de noms propres de la langue visée, comme les prénoms ou les toponymes. Lexis possède actuellement plus de 24 000 noms de personnes.

Repérer dynamiquement de nouvelles entités par apprentissage – Une limite de la majorité des systèmes à base de règles est qu'ils ne peuvent s'adapter automatiquement à un nouveau corpus. Nous proposons une méthode d'apprentissage utilisant les éléments déjà trouvés et les règles de la grammaire de reconnaissance des noms propres pour étendre la couverture du système initial. La méthode est proche de celle proposée par Cucchiarelli et al. (CLV98). Il s'agit d'un cas d'apprentissage à base d'explication (Mit97). Le mécanisme se fonde sur les règles de la grammaire ayant mis en jeu des mots inconnus. Par exemple, la grammaire peut reconnaître *Mr Kassianov* comme étant un nom de personne même si *Kassianov* est un mot inconnu. Les occurrences isolées du mot peuvent dès lors être étiquetées comme nom de personne. L'apprentissage est ici utilisé comme un mécanisme inductif utilisant les connaissances du système (les règles de la grammaire) et les entités préalablement retrouvées (le jeu d'exemples positifs) pour améliorer les performances (on constate un gain de performance de l'ordre de 10 à 15% suivant le texte, soit 0,66 à 0,70 P&R).

Utiliser les structures de discours – Les structures de discours sont une autre source pour l'acquisition de connaissances (Per98). Pour la terminologie, B. Daille et ses collègues ont montré que de nouveaux termes pouvaient être repérés par l'analyse de séquences de textes particulières (DHJR96). Le même principe peut être utilisé pour l'acquisition automatique de nouvelles entités. Nous nous sommes particulièrement intéressés aux énumérations, largement répandues dans les différents types de corpus que nous avons anlysés. Les énumérations sont

facilement repérables par la présence d'un certain nombre de noms de personnes, séparés par des connecteurs (virgules, conjonction, etc.). Par exemple, dans la séquence : <PERSON_NAME> Kassianov </PERSON_NAME> , <UNKNOWN> Kostine </UNKNOWN> **and** <PERSON_NAME> Primakov </PERSON_NAME> où *Kostine* est un mot inconnu, le système peut inférer du contexte que le mot *Kostine* réfère à un nom de personne. Cette règles n'est bien entendu valable que pour les énumérations où apparaissent des noms de personnes et des mots inconnus, mais ne peut pas s'appliquer pour tout type d'énumération. Grâce à cette stratégie, le score de Lexis sur le corpus Communication a atteint 0.84 P&R.

4.2 Gestion des conflits entre stratégies d'étiquetage

Ces processus d'apprentissage peuvent conduire à des conflits de type. C'est le cas, par exemple, quand un mot enregistré comme nom de lieu dans le dictionnaire apparaît comme nom de personne dans une occurrence non ambiguë du texte. Considérons le passage suivant issu du corpus MUC-6 :

```
<SO> WALL STREET JOURNAL (J), PAGE A2 </SO>
<DATELINE> WASHINGTON </DATELINE>
<TXT>
<p> Consuela Washington, a longtime House staffer and an expert in
securities laws, is a leading candidate to be chairwoman of the Securities
and Exchange Commission in the Clinton administration. </p>
```

Il est clair que dans ce texte *Consuela Washington* désigne une personne, mais la première occurrence du mot *Washington* est plus problématique, dans la mesure où la seule information permettant de faire un choix dans la phrase est une connaissance sur le monde. Encore faut-il relativiser cet indice, dans la mesure où, dans l'absolu, un emploi métaphorique ne serait pas à exclure, comme dans *La France veut continuer de diriger le FMI*. En définitive, c'est une analyse des référents qui permet à l'humain de désambiguïser le texte. Un système automatique a peu de chances de s'en sortir. Pour circonscrire ce type de problème, nous proposons de limiter le processus de typage dynamique en cas de conflit au texte en cours d'analyse et non au corpus dans son entier. Cette approche est adaptée au fait que l'on analyse des textes courts, du type des dépêches AFP. Le système va étiqueter toutes les occurrences isolées de *Washington* comme nom de personne dans le texte précédent, mais dans le texte suivant, si une occurrence isolée du mot *Washington* apparaît, le système l'étiquettera comme nom de lieu, selon le dictionnaire. Lorsque plus d'une étiquette a été trouvée de façon dynamique dans un même texte, la première est alors arbitrairement affectée.

5 Analyse des erreurs restantes

Les trois stratégies présentées plus tôt ont permis au système Lexis de passer de 0,50 à 0,84 P&R sur le corpus Communication. Ces performances restent toutefois inférieures à celles enregistrées sur le corpus MUC-6. Parmi les erreurs restantes, nous distinguons les erreurs non résolues (celles qui auraient dû être prises en compte par les mécanismes mis en place) d'autres erreurs que nous n'avons pas prises en compte. Parmi les erreurs non résolues, il reste :

L'incomplétude de la grammaire ou du dictionnaire empêche de reconnaître la séquence *Lloyd Bentsen*, du fait que ni *Lloyd*, ni *Bentsen* ne figuraient dans les dictionnaires.

Les transformations ayant échappé à l'analyse empêche de reconnaître *Robert S. "Steve" Miller* comme une extension de *Robert S. Miller* car ce type de séquence se révèle à la fois trop complexe et trop particulière pour l'analyseur. Étendre la grammaire à ce type de séquence

risquerait d'introduire plus de bruit que de séquences valides.

Les mots fortement ambigus comme *Sun* dans *Sun Tzu*, qui désigne un nom de personne. Si *Tzu* est un mot inconnu et s'il n'y a pas un contexte clair qui permet de désambiguïser la séquence, il est difficile de la repérer de manière correcte.

Les séquences ambiguës, notamment les cas où un nom d'entreprise porte le nom d'une personne. Ainsi, un nom d'entreprise comme *Mary Kay* a été reconnu comme nom de personne.

6 Conclusion

Alors que l'utilisation de règles de grammaire pour l'identification de noms propres dans des textes journalistiques est une tâche bien maîtrisée et donne des résultats comparables aux humains, l'identification de noms propres dans des textes issus d'échanges informels a reçu nettement moins d'attention et est loin d'être maîtrisée. Deux expériences sur des corpus d'échanges informels ont fait chuter les résultats de systèmes performants vers des scores assez bas. Nous avons donc proposé des stratégies d'adaptation basés sur une couverture plus large des dictionnaires, le repérage dynamique et l'utilisation de structures de discours. L'implantation de ces stratégies dans Lexis a fait remonter les scores sur le corpus Communication de 0,50 à 0,84 P&R.

Remerciements Nous tenons à remercier les rapporteurs pour leurs commentaires et M. Lamontagne, du CRDV.

Références

[ABD+95] J. Aberdeen et al. MITRE : Description of the Alembic System as Used for MUC6. Dans (MUC95).

[AHB+95] D. Appelt et al. SRI International FASTUS system : MUC-6 test results and analysis. Dans (MUC95).

[CLV98] A. Cucchiarelli, D. Luzi et P. Velardi. Automatic semantic tagging of unknown proper names. Dans *Actes ACL-COLING 1998*, pp 286-292, Montréal, Canada.

[AM99] D. Appelt et D. Martin. Named Entity Recognition in Speech : Approach and results using the TextPro System. Dans *Proceedings of the DARPA Broadcast News Workshop*. Herndon, Virginia, 1999.

[DHJR96] B. Daille, B. Habert, C. Jacquemin et J. Royaute. Empirical observation of term variations and principles for their description. *Terminology*, 3(2) :197–258, 1996.

[KL98] L. Kosseim et G. Lapalme. Exibum : Un système expérimental d'extraction d'information bilingue. Dans *Actes de la Rencontre Internationale sur l'extraction, le Filtrage et le Résumé Automatique (RIFRA-98)*, pp 129–140, Sfax, Tunisie, 1998.

[KSSW98] F. Kubala, R. Schwartz, R. Stone et R. Weischedel. Named Entity Extraction from Speech. Dans *Proceedings of the DARPA Broadcast News Workshop*. Herndon, Virginia, 1998.

[Mit97] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[MUC95] *Proceedings of the* 6^{th} *Message Understanding Conference*, 1995. Morgan Kaufmann.

[MUC98] *Proceedings of the* 7^{th} *Message Understanding Conference*,
http ://www.itl.nist.gov/iad/894.02/related_ projects/muc/proceedings/muc_ 7_ toc.html.

[Per98] M.-P. Péry-Wooodley. Textual signalling in written texts : a corpus-based approach. Dans *Actes ACL-COLING 1998, workshop Discourse relations and discourse markers*, Montréal, Canada.

[Poi99] T. Poibeau. Le repérage des entités nommées : un enjeu pour les systèmes de veille. Dans *Terminologies Nouvelles (Actes du colloque Terminologie et Intelligence Artificielle, TIA'99, Nantes)*, n. 19, pp 43–51, Nantes, France, 1999.