Rapport technique Projet Gramlab : livrable SP5.1 Use Case Eptica/Lingway : identification d'amorces de reprises

Alain Couillault, Axelle Vinckx, Hugues de Mazancourt, Fanny Grandry, Gaëlle Recourcé

Rapport technique Projet GramLab
Livrable SP 5.1 Use case Eptica/lingway : identification d'amorce de reprise

Alain Couillault, Axelle Vinckx, Hugues de Mazancourt, Fanny Grandry, Gaëlle Recourcé

1. Motivations

Avec une prévision pour 2015 de 4,1 milliards de comptes de messagerie électronique [1], le développement de solutions d'aide au traitement des courriels est devenu un réel enjeu industriel. Pour la plupart, ces traitement nécessitent de séparer le nouveau contenu des reprises de messages précédents. Cette segmentation doit prendre en compte un large éventail de situations, dues à la variété des langues, aux comportements des clients de messagerie, aux actions effectuées par l'utilisateur et à ses comportements épistolaires. Elle peut être effectuée par des grammaires locales de reconnaissance des *amorces de reprises* , suites de caractères qui marquent les frontières entre le nouveau contenu et les reprises. Le temps, donc le coût, de développement de telles grammaires est susceptible d'être diminué par l'induction automatique de règles à partir d'exemples.

Nous avons dans cette expérience évalué l'apport qualitatif d'un générateur automatique de graphes pour la création de grammaires locales de reconnaissance des *amorces de reprise*. Après un rappel sur la segmentation de courriels, nous présentons d'abord l'outil Auto-graph d'induction automatique de grammaires locales, développé dans le cadre du projet GramLab, les corpus utilisés pour cette évaluation, le protocole mis en place et enfin les résultats obtenus.

2. Segmentation d'emails

La définition des zones de courriel varie selon les besoins applicatifs en vue : extraction automatique des coordonnées d'un contact (Laval *et al.* 2009), reconnaissance de signatures de réponses (Carvalho et Cohen 2004), profilage et reconnaissance des auteurs (Estival *et al.* 2007) ou l'identification d'actes de langage (Cohen *et al.* 2004).

(Estival *et al.* 2007) proposent cinq catégories de zones (*Author text*, *signature*, *advertisement*, *quoted text* et *reply lines*) que (Lampert *et al.* 2009) reprennent et étendent à 9 zones de granularité plus fine. Ces derniers ont expérimenté deux approches statistiques pour segmenter les courriels en zones, basées sur un premier découpage grossier du contenu : en une suite de lignes non-vides pour l'une, sur la base d'heuristiques pour l'autre - par exemple, une ligne vide, une ligne de caractères espace ou une ligne commençant par une ponctuation répétée 3 fois ou plus sont considérées comme des frontières de zone- . Le module statistique est ensuite chargé de la

1. Source : (Radicati S 2011)

classification de ces éléments afin de leur attribuer une zone.
(Carvalho et Cohen 2004) classent les lignes d'un courriel en zones sur la base d'heuristiques qui mêlent frontières et contenu. Par exemple :
– une suite de trois lignes vides (*frontière*),
– la ligne courant contient une adresse de courriel reconnue par un patron (*contenu*).
(Estival *et al.* 2007) utilisent une approche semblable ainsi qu'une annotation manuelle. Du fait de son coût, l'annotation manuelle n'est utilisée que pour l'évaluation.
Notre expérimentation concerne le développement d'une grammaire de reconnaissance des amorces de reprise. Une telle grammaire présente l'avantage qu'elle peut être intégrée dans une chaîne de traitements (Laval *et al.* 2009) ou servir à amorcer l'annotation manuelle d'un corpus à des fins d'entrainement et d'en augmenter la cohérence .(Fort et Sagot 2010) ont ainsi montré qu'une pré-annotation automatique diminue les coûts en temps humain d'une campagne d'annotation.

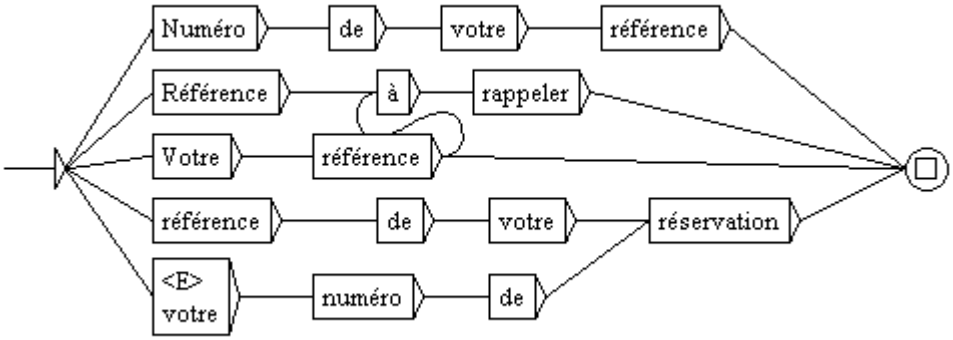
3. Auto-graph

Le projet GramLab [2] a pour objectif de fournir un environnement OpenSource pour le développement et la mise en production de modules d'analyse linguistique. Il propose des outils pour la constitution assistée de corpus (*GramLab Corpus Manager*), la rédaction de grammaires locales (*GramLab IDELing*), et l'intégration de ces grammaires dans une chaîne de traitements (*GramLab Annotator*). GramLab IDELing est un environnement de génie logiciel, basé sur Unitex (Paumier 2006), dédié à la création et à la maintenance de grammaires. Il propose notamment des fonctions de versionnement de fichiers source (CVS) et d'aide au déploiement (Maven).
GramLab IDELing intègre également un module de génération automatique de graphes (Autograph) dont l'entrée est un ensemble d'exemples et la sortie un graphe induit à partir de ces exemples.
Ainsi, la séquence :

```
votre numéro de réservation
numéro de réservation
référence de votre réservation
Votre référence
Référence à rappeler
```

2. http ://www.gramlab.org Le projet GramLab, labellisé par le pôle de compétitivité Cap Digital a été développé dans le cadre du programme FEDER

génère automatiquement le graphe :



Le fichier d'entrée fourni à Auto-graph peut être préparé et contenir, par exemple, des pré-terminaux afin de générer une grammaire plus générique.
Si ce mécanisme d'induction automatique produit rapidement une grammaire locale, le risque de foisonnement des règles et la complexité du graphe résultant peut rendre la modification et la maintenance de ces grammaires difficiles et coûteuses, d'autant qu'Auto-graph permet d'itérer le processus pour enrichir un graphe avec de nouveaux exemples.

4. Les corpus

L'équipe Eptica/Lingway a rassemblé un corpus de courriels volumineux de contenus (courriels professionnels, privés, français, anglais) et de formes (différents usages et clients de messagerie) variés. Quelques centaines de courriels ont été extraits pour constituer un *corpus source* et 54 courriels, qui ont été anonymisés et rassemblés pour constituer un *corpus de test*. Tous les courriels du corpus de test et du corpus source contiennent des amorces de reprise. Le corpus de test est disponible en ligne [3], la documentation et les conditions d'utilisation sont décrits dans la *Charte Ethique et Big Data* (Couillault et Fort 2013) associée .

5. Le protocole

Il s'agit d'obtenir rapidement une grammaire locale qui annote les courriels en y ajoutant un marqueur d'amorce de reprise [4] comme dans l'exemple ci-dessous, où `<ReplyForwardTrigger>` est la marque de début et de fin d'une amorce de reprise [5] :

```
[...]le champ des possibles de l'architecture", a salué Lord Palumbo, membre du jury.
{S}<ReplyForwardTrigger>Harry Potter <harry.potter@gmail.com> a écrit :<ReplyForwardTrigger>
{S}Le Portrait d'une dame (photo) peint par Picasso en 1928 (gouache, encre et
collage sur papier), n'a jamais été exposé et n'était connu que par une mention [...]
```

3. http ://www.gramlab.org/downloads/wp5/CorpusDeTest.rar
4. le texte du message et les noms et adresses de courriel ont été substitués lors de l'anonymisation.
5. `{S}` est ici une marque de séparation de phrases.

Le protocole d'expérimentation a consisté à (i) ébaucher une première version d'une `grammaire locale de travail` (ii) identifier manuellement des patrons génériques d'amorces de reprise (iii) appliquer Auto-graph à ces patrons pour obtenir automatiquement une `grammaire locale induite` (iv) extraire de nouvelles instances d'amorces de reprises avec cette grammaire (v) compléter la grammaire de travail.

5.1. Grammaire de travail

GramLab IDELing fournit en standard une grammaire de segmentation en phrases qui sur-segmente les courriels. L'adaptation de cette grammaire a été effectuée simplement en ajoutant les abréviations courantes de mois.

5.2. Patrons génériques

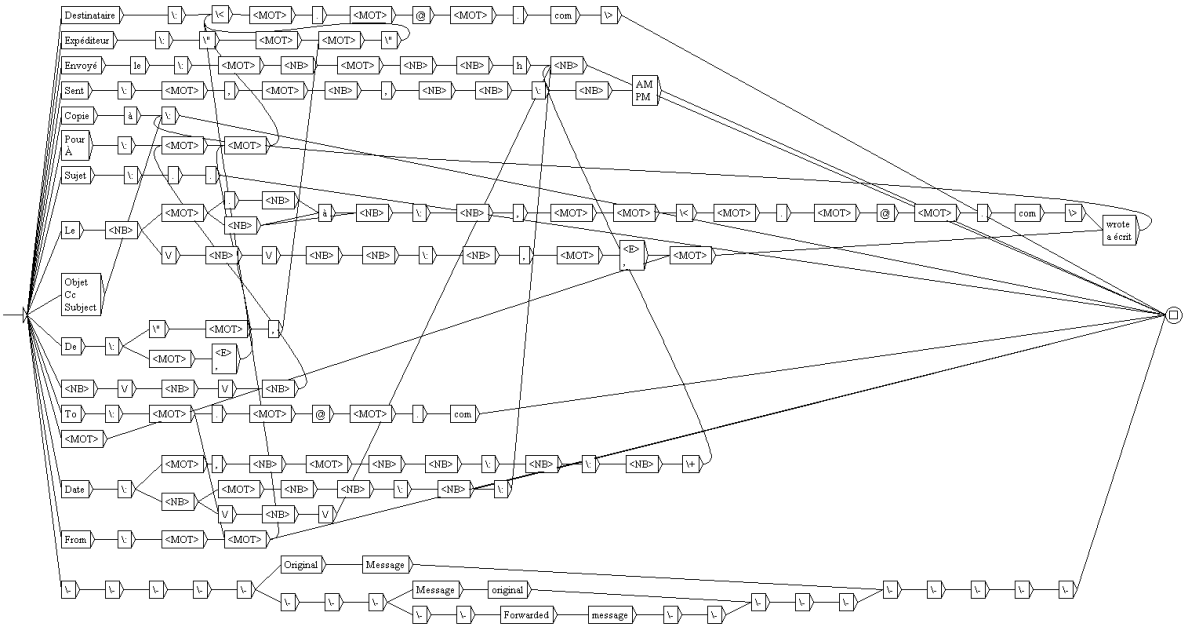
Un repérage manuel dans le corpus de développement a fait ressortir une vingtaine d'instances d'amorces de reprise, qui ont été généralisées en remplaçant les mots non-pertinents par <MOT> et les nombre par <NOMBRE>.
Par exemple :

```
Le 01/07/1976 00:00, Pan, Peter a écrit :
---------- Forwarded message ----------
From: Peter Pan <peter.pan@neverland.com>
Date: 1976/07/01
Subject:
To: Peter Pan <peter.pan@neverland.com>
Cc:
```

a été remplacé par

```
Le <NB>/<NB>/<NB> <NB>:<NB>, <MOT>, <MOT> wrote :
---------- Forwarded message ----------
From: <MOT> <MOT> \<<MOT>.<MOT>@<MOT>.com\>
Date: <NB>/<NB>/<NB>
Subject:
To: <MOT> <MOT> \<<MOT>.<MOT>@<MOT>.com\>
Cc:
```

A partir de ces patrons, AutoGraph a produit une grammaire locale induite :



qui, appliquée au corpus source, a permis d'identifier de nouveaux patrons et d'enrichir la grammaire de développement manuellement via l'interface graphique de GramLab IDELing.

6. Evaluation

La grammaire résultante a ensuite été appliquée au corpus de test pour annoter les empans correspondant aux amorces de reprises :

```
<ReplyForwardTrigger>Date : 01 juillet 1976 00:00:00<ReplyForwardTrigger>
<ReplyForwardTrigger>À : Peter Pan <peter.pan@neverland.com><ReplyForwardTrigger>
<ReplyForwardTrigger>Le 01/07/1976 00:00, Peter Pan a écrit :<ReplyForwardTrigger>
<ReplyForwardTrigger>________ Message original ________<ReplyForwardTrigger>
<ReplyForwardTrigger>> Sujet:<ReplyForwardTrigger>
<ReplyForwardTrigger>Date : Sat, 01 juill 1976 00:00:00 +0100<ReplyForwardTrigger>
<ReplyForwardTrigger>De : Pan, Peter <peter.pan@neverland.com><ReplyForwardTrigger>
<ReplyForwardTrigger>Pour : Peter Pan <peter.pan@neverland.com><ReplyForwardTrigger>
```

Deux types de mesures ont été effectuées :

- évaluer le nombre d'*amorces de reprises reconnus*,
- évaluer le nombre de *courriels bien segmentés*.
  On considère qu'un courriel est bien segmenté lorsqu'aucune ligne de la reprise ne se trouve dans le texte de l'auteur, c'est à dire, généralement, lorsque la première instance d'amorce de reprise est identifiée.

Nous disposions également d'un module de segmentation de la société Kwaga comprenant une cinquantaine d'expressions régulières développé au long cours sur plusieurs milliers de cour-

riels. Nous avons également appliqué les deux grammaires sur un extrait du corpus Enron (Klimt et Yang 2004), mais l'homogénéité de ce corpus n'a pas permis d'aboutir à une évaluation pertinente. Le tableau suivant donne les précisions, rappels et f-mesures pour la grammaire GramLab et le module Kwaga, pour la reconnaissance des amorces de reprise (amr) et la segmentation en zones.

	amr		segmentation	
	gramlab	kwaga	gramlab	kwaga
précision	100,00%	100,00%	96,30%	92,59%
rappel	97,92%	88,19%	96,30%	92,59%
F-mesure	98,95%	93,73%	96,30%	92,59%

Le rappel est sensiblement meilleur pour les deux tâches avec la grammaire développée dans l'environnement GramLab. Le volume et la nature du corpus utilisé pour le développement du module Kwaga et l'hétérogénéité du corpus de test diminuent un éventuel biais lié au fait que le corpus de test et la grammaire ont été constitués à partir de sources proches.

Les deux approches (GramLab et Kwaga) sont basées sur une déclaration explicite des segments à reconnaître, qui sont suffisamment non ambigus pour obtenir une précision maximale. Ces résultats sont à comparer avec ceux obtenus par des approches statistiques pour des tâches légèrement différentes : (Estival *et al.* 2007) obtiennent une précision de 0,987 pour la reconnaissance des zones de réponse et (Lampert *et al.* 2009), qui visent à catégoriser toutes les lignes d'un courriel, obtiennent pour la reconnaissance de la zone auteur [6] une précision de 0.868.

Par ailleurs, tous les courriels du corpus de test sont annotés, et une seule marque est attendue par courriel pour la segmentation en zones. La précision, le rappel et la f-mesure sont donc identiques pour cette tâche.

Les courriels du corpus de test contiennent chacun entre 1 et 5 amorces de reprise, le tableau ci-dessous affine les résultats précédents et donne une mesure de bruit plus précise : seule l'amorce de reprise qui sépare effectivement la zone auteur du reste du courriel est considérée comme pertinente [7].

			rappel		précision
nb d'amr par mail	fréquence	Gramlab	kwaga	Gramlab	Kwaga
1	27	100%	100%	100%	100%
2	0				
3	2	0%	100%	0%	50%
4	64	100%	87,5%	25%	21,88%
5	48	90%	80%	18,75%	16,67%

La méthode de segmentation employée favorise le rappel à la précision, car elle vise à recon-

6. c'est dire, dans leur cas, le texte rédigé par l'émetteur, hors textes automatiques tels que la signature ou le *disclaimer*.

7. Aucun courriel ne contient deux amorces de reprise, les résultats ne sont donc pas donnés pour cette catégorie

naître tous les marqueurs de frontières, quelque soit leur contexte, et laisse ainsi la responsabilité du découpage effectif à la suite du traitement. Cependant, l'augmentation simultanée de la précision et du rappel pour les catégories des courriels contenant cinq amorces de reprise laisse penser que le protocole mis en place permet une meilleur identification de ces marqueurs, grâce aux exemples collectés, et donc une meilleure grammaire résultante.

7. Conclusion et perspectives

Nous avons montré que pour une tâche simple - la reconnaissance des amorce de reprise-, l'induction automatique à partir d'exemples permet à la fois de réduire les coût de développement et d'augmenter la qualité d'une grammaire locale. Cette approche est également applicable à la maintenance des grammaires, pour diminuer le coût de leur adaptation.
Le protocole pourrait être appliqué à la reconnaissance de phénomènes plus complexes, qui pourraient nécessiter une itération sur l'induction automatique de grammaire. Une telle expérience permettrait d'évaluer l'influence du choix des patrons fournis à Auto-graph, et d'évaluer les impacts sur les coûts de développement et de maintenance et sur la qualité et la cohérence des grammaires résultantes.

Remerciements

La grammaire a été développée par Axelle Vinckx, Eptica/Lingway

Bibliographie

Carvalho V. R. & Cohen W. W. (2004), « Learning to Extract Signature and Reply Lines from Email », in : *CEAS 2004 - First Conference on Email and Anti-Spam*, Mountain View, CA.

Cohen W. W., Carvalho V. R. & Mitchell T. M. (2004), « Learning to Classify Email into "Speech Acts" », in : Lin D. & Wu D. (Eds), *Proceedings of EMNLP 2004* : Association for Computational Linguistics, Barcelona, Spain : 309–316.

Couillault A. & Fort K. (2013), *Charte Ethique et Big Data : parce que mon corpus le vaut bien*, (en cours de soumission).

Estival D., Gaustad T., Pham S. B., Radford W. & Hutchinson B. (2007), « Author profiling for English emails », in : *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, Melbourne, Australia : 263–272.

Fort K. & Sagot B. (2010), « Influence of Pre-annotation on POS-tagged Corpus Development », in : *Fourth ACL Linguistic Annotation Workshop*, Uppsala : 56–63.

Klimt B. & Yang Y. (2004), « Introducing the Enron corpus », in : *First Conference on Email and Anti-Spam (CEAS)*.

Lampert A., Dale R. & Paris C. (2009), « Segmenting email message text into zones », in : *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2 - Volume 2*, (EMNLP '09) : Association for Computational Linguistics, Stroudsburg, PA, USA : 919–928.

Laval P., Meunier F., Recourcé G. & Surcin S. (2009), *Kwaga : une chaîne UIMA d'analyse de contenu des mails - Proposition de démonstration*.

Paumier S. (2006), *Unitex - Manuel d'utilisation*.

Radicati S H. Q. (2011), *Email Statistics Report, 2011-2015*, http ://www.radicati.com/wp/wp-content/uploads/2011/05/Email-Statistics-Report-2011-2015-Executive-Summary.pdf, The Radicati Group, Inc., last checked 3/30/2013.