Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations

Yayoi Nakamura-Delloye[1] Eric Villemonte de La Clergerie[1]
(1) ALPAGE, INRIA-Rocquencourt, Domaine de Voluceau Rocquencourt B.P.105
78153 Le Chesnay
yayoi@yayoi.fr, eric.de_la_clergerie@inria.fr

Résumé. Le présent article décrit un travail en cours sur l'acquisition des patrons de relations entre entités nommées à partir de résultats d'analyse syntaxique. Sans aucun patron prédéfini, notre méthode fournit des chemins syntaxiques susceptibles de représenter une relation donnée à partir de quelques exemples de couples d'entités nommées entretenant la relation en question.

Abstract. This paper describes our current work on the acquisition of named entity relation patterns from parsing results. Without any predefined pattern, our method provides candidate syntactic paths that represent a given relationship with a small seed set of named entity pairs on this relationship.

Mots-clés : Extraction des connaissances, extraction des patrons, relation des entités nommées, arbre syntaxique dépendanciel.

Keywords: Knowledge extraction, pattern extraction, named entity relation, syntactic dependency tree.

1 Introduction

Le présent article décrit des travaux en cours de réalisation dans le cadre du projet SCRIBO [1] ayant pour objectif la mise au point d'algorithmes et d'outils collaboratifs pour l'extraction de connaissances à partir de textes et d'images. Dans ce projet, l'extraction de connaissances est considérée comme un processus cumulatif commençant par des traitements de corpus en amont pour faire émerger dans le corpus syntaxiquement analysé des régularités traduisant des informations sémantiques susceptibles de trouver place au sein d'une ontologie. L'arbre syntaxique comprenant plus d'information sur les relations entre les éléments de la phrase, il existe déjà des travaux recourant à cette représentation pour l'extraction de connaissances (Kramdi *et al.*, 2009). Nos travaux s'intéressent plus particulièrement aux relations entre les entités nommées (EN ci-après). Nous avons donc décidé d'examiner d'abord les chemins entre deux EN existant dans un arbre syntaxique de dépendance, pour examiner leurs relations et trouver une méthode d'extraction des patrons à partir d'arbres syntaxiques.

L'état actuel de nos travaux résulte de trois études réalisées successivement, chacune ayant fourni des hypothèses pour la suivante. La première étude (§ 2) a consisté à extraire les chemins entre tous les couples

1. Le dossier SCRIBO labellisé par le pôle de compétitivité System@tic et soumis au Fonds Unique Interministériel (FUI) a été accepté par la DGE pour financement.

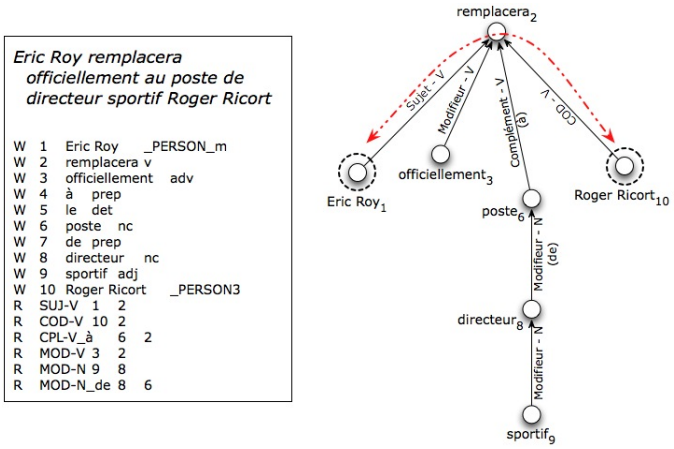

FIGURE 1 – Résultat d'analyse syntaxique et arbre de dépendance

de deux entités nommées, présents dans un arbre de dépendance résultant d'une analyse syntaxique. La seconde (§ 3) a porté sur le regroupement des chemins et des EN. La troisième et dernière (§ 4) a tenté de mettre en correspondance les chemins extraits et les relations entre EN. Nous présenterons également quelques résultats de notre première expérience (§ 5) avant de terminer avec les perspectives de nos travaux.

2 Extraction des chemins syntaxiques de relation

Le fichier d'entrée initial contient un résultat d'analyse syntaxique avec étiquetage des EN, fourni par l'analyseur FRMG [2], non vérifié. Étant donné qu'il contient beaucoup d'informations de diverses natures, on en extrait d'abord les seules données nécessaires à la construction de l'arbre syntaxique et celles sur les constituants de la phrase correspondant aux nœuds de l'arbre. Avec les données extraites du résultat d'analyse syntaxique, l'arbre syntaxique d'entrée est construit à partir de l'ensemble des relations de dépendance entre les constituants (cf. Figure 1).

Notre première hypothèse a été, comme dans (Bunescu & Mooney, 2005), que la relation entre deux EN était représentée dans l'arbre syntaxique par le chemin reliant les deux nœuds leur correspondant. Ainsi, dans l'arbre de la figure 1, la relation entre les deux ENs, *Eric Roy* et *Roger Ricort*, est représentée par le chemin reliant leurs nœuds tracé par la ligne non contigue, constitué en une suite d'arcs et de nœuds intermédiaires : `=(Sujet-V)=> remplacera <=(COD-V)=`. La première étude a consisté en l'extraction de tous ces chemins entre deux EN, que nous appelons « chemins syntaxique de relation ». L'extraction des chemins entre deux EN revient à la recherche du chemin entre les deux nœuds représentant une EN dans cet arbre. Cette opération peut être réalisée par un algorithme classique de recherche du plus court chemin. Nous avons adopté l'algorithme de Floyd-Warshall avec la transformation de l'arbre syntaxique d'entrée en un graphe symétrique.

2. Pour la description de l'ensemble des analyseurs utilisés et leurs notations, voir notamment (Villemonte de La Clergerie *et al.*, 2009) et (Villemonte de La Clergerie, 2010).

3 Constitution des ensembles de chemins et de couples d'EN

Les chemins ainsi obtenus présentent une diversité considérable. Nous nous sommes donc limités dans un premier temps aux EN typées « organisation » ou « individu ». La seconde étude a donc eu pour objectif de trouver un moyen de filtrage pour cet ensemble encore confus, et a consisté en un regroupement des chemins partageant les mêmes EN de départ et de fin, ainsi que celui des couples d'EN partageant le même chemin de relation, afin d'examiner leur fréquence.

Ainsi, nous pouvons obtenir un ensemble d'ensembles de chemins (corrects et erronés confondus) qui partagent les mêmes EN de départ et de fin, telles que, respectivement `Ali Bongo` et `André Mba Obame` dans l'exemple suivant [3] :

```
> Ensemble 1 : Ali Bongo (individu) - André Mba Obame (individu)
==>devance<==ministre<==
==>a==>remporté<==élection<==tour<==30 août<==soit==>devant<==ministre<==
==>a==>remporté<==élection<==tenue<==voix<==
...
```

Nous obtenons également un ensemble d'ensembles de couples d'EN (corrects et erronés confondus) qui partagent le même chemin de relation. L'exemple suivant est l'ensemble constitué des couples des $\texttt{EN}_1$, $\texttt{EN}_2$ partageant le même chemin de relation `==> (MOD-N)` qui signifie que $\texttt{EN}_1$ dépend de $\texttt{EN}_2$ et qu'ils sont en relation `Modifieur-Nom`.

```
> Ensemble 1 : ==> (MOD-N)
Gilles Carrez (individual) - UMP (organization)
Sordo (individual) - Dani (individual)
Real (individual) - Madrid (organization)
Ban Ki-moon (individual) - l'ONU (organization)
...
```

4 Mise en correspondance des chemins et des relations des EN

4.1 Hypothèses pour la mise en correspondance

L'examen des résultats de ce regroupement nous a permis de poser l'hypothèse suivante :

> **Hypothèse :** Si le couple pen_0 constitué de $en_1^{type_s}$ et $en_2^{type_t}$, entretient la relation $\texttt{R}_0$, alors tous les chemins chm_i qui relient ce couple représentent la même relation, $\texttt{R}_0$, et tous les couples, pen_j, mis en relation via un de ces chemins entretiennent également cette relation $\texttt{R}_0$.
>
> Soit E_{pen_n}, ensemble des chemins, chm_i, reliant les EN constituant pen_n
> si : $f_{rel}(pen_n^{type_x,type_y}) = R_s$
> alors : $\forall pen_j^{type_x,type_y} \in E_{chm_j}$ tel que $chm_j \in E_{pen_n}, f_{rel}(pen_j) = R_s$

Prenons un exemple avec la relation d'appartenance. Supposons que nous sachions que « Gilles Carrez » (individu) **appartient à** l'organisation « UMP ». Étant donné que la paire `(Gilles Carrez`$_{individu}$`,`

3. Les EN et les arcs de chemin représentant une relation de dépendance sont typés mais ils ne sont pas toujours explicitement marqués dans les exemples cités.

UMP$_{organisation}$) est reliée par le chemin ==> (MOD−N), nous considérons que toutes les paires du type (individu A, organisation B) appartenant à l'ensemble des couples reliés par ce chemin sont des paires d'EN en relation d'appartenance tel que « individu A » appartient à « organisation B » et que ce chemin ==> (MOD−N) est un patron pour la relation d'appartenance.

4.2 Méthode de définition automatique des chemins de relation

La méthode d'extraction automatique des chemins de relation basée sur cette hypothèse se fonde sur un principe d'« induction » utilisé dans les travaux d'identification des patrons textuels, tels que ceux de (Hearst, 1992), et consiste à, pour une relation donnée notée R, donner au système quelques exemples de couples des EN en relation R afin qu'il nous fournisse en retour tous les chemins représentant cette relation R qu'il a trouvés dans le corpus (cf. Fig. 2). Cette méthode a déjà été utilisée dans des travaux sur l'extraction des patrons de relations des EN (Brin, 1998), (Agichtein & Gravano, 2000), mais notre approche se distingue des leurs par l'exploitation des chemins des arbres syntaxiques. Ces chemins de relation pourraient ensuite être transformés en expressions textuelles afin de fournir les patrons « classiques » utilisés pour l'extraction des relations. Avec cette méthode, très peu de connaissances préalables sont requises (quelques exemples de couples d'EN pour une relation donnée) et les chemins longs, dus très souvent à une analyse syntaxique erronée, n'ont pas d'influence sur le résultat final puisqu'ils ne sont jamais partagés par plusieurs paires d'EN.

5 Expérience pour la relation d'« appartenance »

Afin d'évaluer la fiabilité des chemins extraits en tant que patrons de relations, nous avons examiné le résultat d'extraction de patrons de la relation d'« appartenance » basée sur notre méthode décrite dans les sections précédentes. L'évaluation consiste à examiner la nature des couples d'EN extraits par les chemins fournis.

Les premières expériences ont été effectuées avec un corpus constitué de dépêches AFP de deux mois, du 20 mai au 29 juin 2009, annoté syntaxiquement sans aucune vérification manuelle. Ce corpus comporte 1 173 173 phrases avec 381 745 entités nommées. À partir de 146 304 phrases du corpus contenant plus d'une EN, nous avons extraits 2 128 chemins reliant un couple d'EN, et 27 225 paires d'EN mises en relation. Pour l'identification de patrons, nous avons préparé 12 exemples de couples des EN (*individual, organization*) en relation d'« appartenance », récupérés manuellement depuis Wikipédia, tels que (*Xavier Bertrand, UMP*), (*Martine Aubry, PS*), (*François Bayrou, MoDem*), (*Marie-George Buffet, PCF*).

Nous avons extrait 137 chemins de relations [4]. Toutefois, seuls 20 chemins étaient productifs – reliant plus d'un couple d'EN –, dont 7 exemples sont présentés ci-dessous avec le texte d'origine à partir duquel ils ont été extraits.

1. ==(MOD−N)=> président(nc) <=(MOD−N(de))==
   « président du *MoDem François Bayrou* »
2. <=(JUXT) == secrétaire(nc)<=(MOD−N(de))==
   « la première secrétaire du *PS*, *Martine Aubry* »

4. L'expérience a été réalisée sur un MacBookPro Core i7 2,66 GHz, 8 Go de RAM, Mac OX 10.6.3. Le temps de calcul pour ce corpus de 147,9 Mo (après le pré-traitement) a été de 6 minutes, après l'optimisation du classement des données par l'utilisation des arbres binaires.

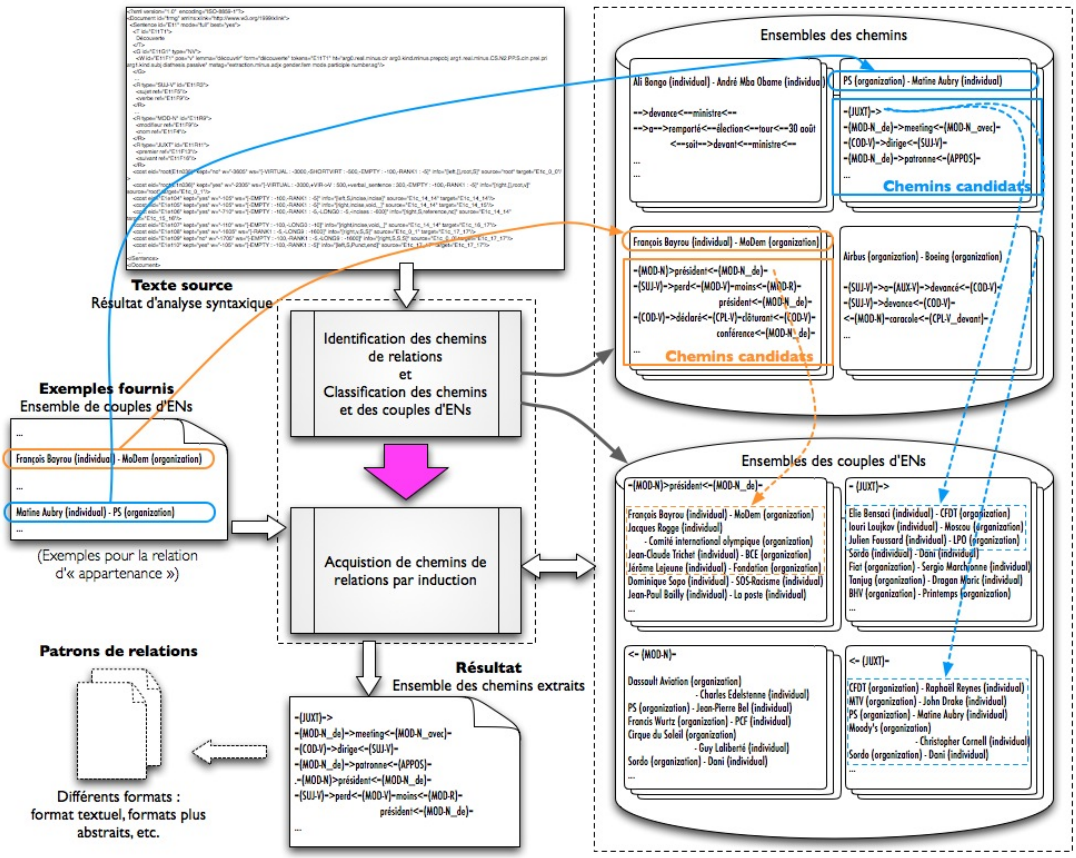

FIGURE 2 – Procédure générale d'extraction des chemins de relations

3. ==(APPOS)=> patron(nc) <=(MOD-N(de))==
« s'est félicité *Xavier Bertrand*, patron de *l'UMP* »
4. <==(MOD-A(pour))
« on relève les noms de ... , de *Marie-George Buffet* pour *le PCF* »
5. ==(SUJ-V)=> dirige(v) <=(COD-V)==
« Pour M. Delors dont la fille *Martine Aubry* dirige *le PS* , ... »
6. ==(SUJ-V)=> demeurait(v) <=(CPL-V(à))== tête(nc) <=(MOD-N(de))==
« la première secrétaire *Martine Aubry* demeurait légimite à la tête du *PS* »
7. <==(MOD-N)était(v)<==(CPL-V(en))position(nc)<==(MOD-N(sur))liste(nc)<==(MOD-N(de))
« Mme *Le Pen*, qui était en deuxième position sur la liste du *FN* à Hénin-Beaumont pour les élections municipales de 2008 »

Avec ces chemins de relations, ont été extraites 1 469 paires d'EN supposées en relation d'appartenance. Après avoir éliminé les erreurs d'étiquetage des EN, nous avons vérifié manuellement 178 paires, et nous avons compté 149 couples corrects contre 29 incorrects. Ce premier résultat nous semble encourageant et les hypothèses méritent de continuer à être vérifiées.

6 Perspectives

Il existe de nombreuses perspectives intéressantes. Il serait sans doute intéressant de réaliser une évaluation plus fine du résultat afin de concevoir un mécanisme de score qui proposerait une indication de fiabilité des chemins (en fonction par exemple de la fréquence). De plus, il faudrait trouver d'autres relations

et analyser les résultats pour évaluer la méthode. Il est également possible d'envisager l'obtention de patrons textuels à partir des chemins représentant les relations intéressantes afin d'améliorer la portabilité des résultats obtenus avec notre méthode, au profit des méthodes classiques basées sur les patrons. Par ailleurs, il est également possible d'envisager de concevoir une méthode non-supervisée par inspiration des travaux existants tels que (Hasegawa *et al.*, 2004) (Zhang *et al.*, 2005) (Chen *et al.*, 2005) (He *et al.*, 2006).

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, p. 85–94.

BRIN S. (1998). Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, p. 172–183.

BUNESCU R. & MOONEY R. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 724–731, Vancouver, British Columbia, Canada : Association for Computational Linguistics.

CHEN J., JI D.-H., TAN C. L. & NIU Z.-Y. (2005). Automatic relation extraction with model order selection and discriminative label identification. In *IJCNLP*, p. 390–401.

HASEGAWA T., SEKINE S. & GRISHMAN R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 415–422, Barcelona, Spain.

HE T., ZHAO J. & LI J. (2006). Discovering relations among named entities by detecting community structure. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, p. 42–48.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, p. 539–545.

KRAMDI S. E., HAEMMERLÉ O. & HERNANDEZ N. (2009). Approche générique pour l'extraction de relations à partir de textes. In *Actes d'IC'09*.

VILLEMONTE DE LA CLERGERIE É. (2010). Convertir des dérivations TAG en dépendances. In *TALN 2010*.

VILLEMONTE DE LA CLERGERIE É., SAGOT B., NICOLAS L. & GUÉNOT M.-L. (2009). FRMG : évolutions d'un analyseur syntaxique tag du français. In *Journée ATALA "Quels analyseurs syntaxiques pour le français ?"*.

ZHANG M., SU J., WANG D., ZHOU G. & TAN C. L. (2005). Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In *IJCNLP*, p. 378–389.