Extraction des chemins entre deux entités nommées en vue de l'acquisition des patrons de relations

Yayoi Nakamura-Delloye

Extraction des chemins entre deux entités nommées en vue de l'acquisition des patrons de relations

Yayoi Nakamura-Delloye

ALPAGE, INRIA-Requencourt
Domaine de Voluceau Rocquencourt B.P.105 78153 Le Chesnay
yayoi@yayoi.fr

Résumé : le présent article décrit un travail en cours sur l'extraction des chemins syntaxiques pour étudier la possibilité d'une acquisition des patrons de relations entre entités nommées.

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

1 Introduction

Le présent article décrit des travaux en cours de réalisation sur l'acquisition de patrons de relations sémantiques. Ces travaux sont menés dans le cadre du projet ANR SCRIBO 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. Nos travaux s'intéressent plus particulièrement aux relations entre les entités nommées (EN ci-après) et nous proposons une méthode d'acquisition de chemins syntaxiques qui pourraient servir aux patrons de relations. Notre méthode d'acquisition se fonde sur deux types d'hypothèses, fournissant chacun la base des deux grandes étapes de la procédure : la classification des chemins et des couples d'ENs, et l'identification de patrons de relations.

Nous allons tout d'abord décrire la tâche d'extraction de relation en passant en revue les études existantes (§ 2), ainsi que nos données, arbres syntaxiques dépendanciels, et le principal élément de travail, le chemin de relation (§ 3). Nous nous intéresserons ensuite aux deux tâches constituant la procédure d'acquisition de patrons : la première étape de classification (§ 4) et la deuxième étape d'acquisition de patrons (§ 5). Nous présenterons également quelques résultats de notre première expérience (§ 6) avant de terminer avec les perspectives de nos travaux.

2 Extraction des relations des entités nommées

L'extraction des relations consiste à identifier différentes relations sémantiques à partir de textes. En partant de l'hypothèse que certains éléments linguistiques peuvent être utilisés pour accéder dans des textes à une relation conceptuelle, ont été proposés plusieurs travaux sur l'extraction des relations basées sur des patrons textuels (Séguéla & Aussenac-Gilles, 1999; Condamines, 2002; Aussenac-Gilles & Jacques, 2008). Nos travaux visent à déterminer non pas les contextes linéaires mais les structures syntaxiques permettant de repérer ces relations conceptuelles afin de proposer des patrons abstraits.

Dans un premier temps, nous nous intéressons notamment aux relations entre deux entités nommés apparaissant dans une phrase. Nous entendons par entités nommées « tous les éléments du langage définis par référence : les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantité » (Friburger, 2006). Dans la phrase « *Martine Aubry* dirige le *PS* », nous avons deux entités nommées, une du type « individu » (*Martine Aubry*) et une du type « organisation » (*PS*). L'extraction de relations consiste alors ici à repérer la relation « est dirigeant(e) de » entre *Martine Aubry* et *PS*. Un certain nombre de méthodes destinées à l'extraction de relations entre ENs ont déjà été proposées (Brin, 1998; Agichtein & Gravano, 2000; Hasegawa *et al.*, 2004; He *et al.*, 2006), mais notre approche se distingue des leurs par l'exploitation non pas des textes linaires, mais des arbres syntaxiques. L'arbre syntaxique comprenant plus d'information sur les relations syntaxiques entre les éléments de la phrase, il existe déjà des travaux recourant à cette représentation pour cette tâche d'extraction (Zhang *et al.*, 2005; Greenwood & Stevenson, 2006; Kramdi *et al.*, 2009). Nos travaux se caractérisent par le fait qu'ils s'intéressent non pas à certaines relations syntaxiques particulières mais aux chemins reliant deux éléments de phrase représentant des ENs dans l'arbre dépendanciel.

3 Chemins syntaxiques de relations

Les données à partir desquelles nous effectuons l'extraction sont des résultats d'analyse syntaxique en dépendance. Cette analyse consiste à fournir comme résultat toutes les relations syntaxiques existant entre les éléments de phrase en terme de dépendance (Tesnière, 1988). À partir de ce résultat, un arbre syntaxique dépendanciel est construit. La figure 1 montre un exemple de résultat d'analyse syntaxique en dépendance et de son arbre. Notre première hypothèse a été, comme dans Bunescu & Mooney (2005), que la relation entre deux ENs était représentée dans l'arbre syntaxique par le chemin le plus court 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 : `Eric Roy =(Sujet-V)=> remplacera <=(COD-V)= Roger Ricort`. L'extraction des chemins de relations revient à la recherche du plus court chemin entre les deux nœuds représentant une EN dans l'arbre. Cette opération peut être réalisée par un algorithme classique conçu à cet effet tel que celui de Floyd. Avant d'appliquer cet algorithme, l'arbre syntaxique d'entrée doit alors être transformé en un graphe non orienté.

L'extraction des chemins de relations consiste ensuite en les deux grandes étapes de

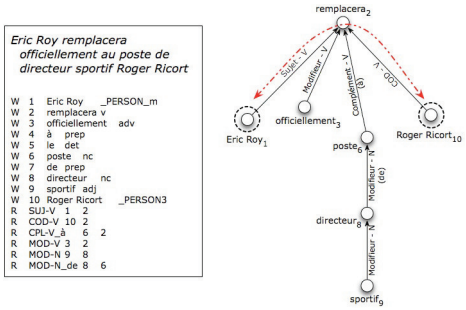

FIGURE 1 – Chemins entre deux ENs dans un arbre dépendanciel

la procédure (cf. Fig. 2) : classification des chemins et des couples d'ENs ; acquisition de patrons de relations.

4 Classification des chemins et des couples d'ENs selon les contextes syntaxiques

Partant de l'hypothèse que l'on pourrait trouver certains traits sémantiques communs entre les éléments qui partagent les mêmes contextes, les couples d'ENs et les chemins sont regroupés selon leurs contextes.

Toutes les paires (pen_i) d'ENs (en_1, en_2) partageant le même chemin de relation (chm_j) sont regroupées, constituant alors un ensemble d'ensembles de paires d'ENs E_{chm_j}. Ainsi, nous obtenons par exemple l'ensemble constitué des couples des ENs partageant le même chemin de relation = `(MOD-N)=>`.

```
> Ensemble 1 : = (MOD-N)=>
Gilles Carrez (individual) - UMP (organization)
Iouri Loujkov (individual) - Moscou (organization)
Sordo (individual) - Dani (individual)
...
```

De même, tous les chemins (chm_s) qui partagent la même paire (pen_t) d'ENs de départ (en_1) et de fin (en_2) sont regroupés, constituant alors un ensemble d'ensembles de chemins E_{pen_t}. Ainsi, nous obtenons par exemple l'ensemble de chemins qui partagent les mêmes ENs de départ et de fin, respectivement `Ali Bongo` et `André Mba Obame` :

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

5 Acquistion de chemins de relations par induction

Une fois ces ensembles de chemins et de couples ENs constitués, la tâche consiste à déterminer la relation sémantique que représentent ces chemins afin de les définir

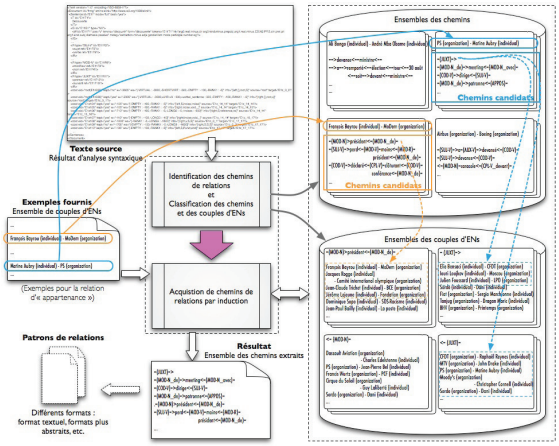

FIGURE 2 – Procédure générale d'extraction des chemins de relations

comme patrons d'une relation donnée.

Pour l'acquisition de patrons de relations, nous avons posé l'hypothèse suivante :

> **Hypothèse :** Si le couple pen_0 constitué de $en_1^{type_s}$ et $en_2^{type_t}$, entretient la relation $\mathtt{R}_0$, alors tous les chemins chm_i qui relient ce couple représentent la même relation, $\mathtt{R}_0$, et tous les couples, pen_j, mis en relation via un de ces chemins entretiennent également cette relation $\mathtt{R}_0$.
>
> Soit E_{pen_n}, ensemble des chemins, chm_i, reliant les EN constiuant 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 $(\mathtt{Gilles\ Carrez}_{individu}, \mathtt{UMP}_{organisation})$ 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'ENs 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.

La méthode d'extraction automatique des chemins de relations basée sur ces hypothèses 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 ENs en relation `R` afin qu'il nous fournisse en retour comme patrons de cette relation tous les chemins qu'il a trouvés dans le corpus. L'avantage de l'utilisation de chemins pour patrons est que c'est un format abstrait. Ils pourraient ensuite être transformés en expressions textuelles afin de fournir les patrons « classiques » utilisés pour l'extraction des relations. De plus, il est également plus aisé d'avancer encore leur niveau d'abstraction que pour les patrons textuels par des transformations de schéma syntaxique.

6 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'ENs extraits par les chemins fournis.

Les premières expériences ont été effectuées avec un corpus constitué de dépêches AFP d'un mois, du 20 mai au 24 juin 2009, annoté syntaxiquement sans aucune vérification manuelle. Ce corpus comporte 26 990 phrases avec 8 515 entités nommées. À partir de ce corpus, nous avons extraits 3 401 chemins reliant un couple d'ENs, et 90 075 paires d'ENs 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 87 chemins de relations, 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* ≫
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 813 paires d'ENs supposées en relation d'appartenance, dont 353 ont été vérifiées manuellement. Nous avons compté 149 couples corrects et 204 incorrects, mais les erreurs de 175 couples sont dues à un mauvais étiquetage des ENs. Ainsi, le système a fourni 149 couples corrects contre 29 incorrects. L'évaluation de résultat doit être réalisée de manière plus fine pour chaque chemin afin de déterminer la nécessité d'un mécanisme de score qui proposerait une indication de fiabilité de chemins. Néanmoins, ce premier résultat nous semble encourageant et les hypothèses méritent de continuer à être vérifiées.

7 Perspectives

Nos travaux, encore en phase de démarrage, contiennent beaucoup d'hypothèses dont nous devrons démontrer la justesse avec des expériences et des analyses de résultats. Toutefois, il existe également de nombreuses perspectives intéressantes. Il serait sans doute intéressant de définir des coûts afin de calculer les scores permettant d'indiquer la pertinence 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) He *et al.* (2006).

Références

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

AUSSENAC-GILLES N. & JACQUES M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with caméléon. *Terminology*, **14**(1), 45–73.

BRIN S. (1998). Extracting patterns and relations from the world wide web. In *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.

CONDAMINES A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*, **8**(1), 141–162.

FRIBURGER N. (2006). Linguistique et reconnaissance automatique des noms propres. *Meta*, **51**(4), 621–847.

GREENWOOD M. A. & STEVENSON M. (2006). Improving semi-supervised acquisition of relation extraction patterns. In *IEBeyondDoc'06 : Proceedings of the Workshop on Information Extraction Beyond The Document*, p. 29–35, Morristown, NJ, USA : Association for Computational Linguistics.

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*.

SÉGUÉLA P. & AUSSENAC-GILLES N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes d'IC'99*.

TESNIÈRE L. (1988). *Éléments de Syntaxe Structurale*. Paris : KLINCKSIECK, deuxième edition. Cinquième tirage.

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.