Vol.2011-DBS-152 No.1
Vol.2011-IFAT-103 No.1
2011/8/2

Named entity extraction for ontology enrichment

Yayoi Nakamura-Delloye[†1]

We propose in this paper two pattern-based named entity extraction methods for ontology enrichment. The proposed methods are characterized by the use of entity relation patterns obtained by our unsupervised extraction method. These patterns correspond to syntactic paths that connect two named entities in dependency trees. This work aims to take advantage of parsing benefits and also offers solutions for parsing disadvantage.

1. Introduction

Knowledge organization using ontological resources is a valuable issue that arouses the interest of researchers in natural language processing (NLP). World knowledge is important in many NLP tasks, such as automatic text understanding. NLP technologies are also used to acquire knowledge from text in order to enrich such resources. This paper presents knowledge extracting methods using NLP technologies for ontological resource enrichment. The proposed methods aim to acquire from large corpora new named entities (NE) such as people or organizations using automatically extracted patterns.

Pattern-based extraction methods are used for relation discovery[7),10)] and also for entity extraction[1),2),4),8)]. Most methods use patterns provided by a semi-supervised approach based on induction using some seeds. The key feature of our approach is the use of syntactic patterns obtained by an unsupervised technique of relation extraction (RE). We also try to show how our approach takes advantage of parsing benefits and offers solutions for parsing disadvantage.

We first describe the framework for ontology enrichment (Section 2). We then present our named entity extraction methods using entity relation patterns (Section 3). Finally, we examine the results of experiments we conducted and discuss the performance of each approach (Section 4).

†1 Institut National de Recherche en Informatique et en Automatique (INRIA), UMR-I 001, ALPAGE, Univ Paris Diderot, Sorbonne Paris Cit, F-75013 Paris, France

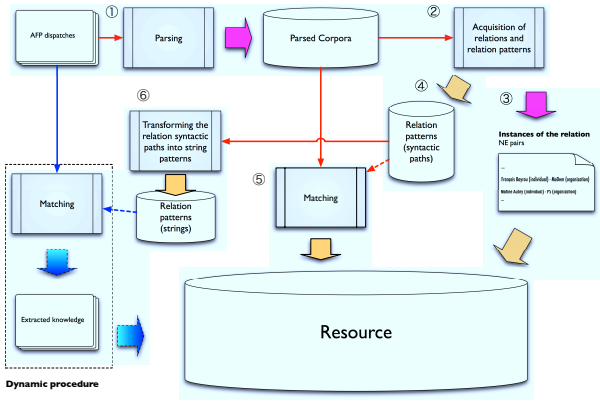

Fig. 1 Framework for ontology enrichment

2. Framework

Within the semantic enrichment experiments of Agence France-Presse (AFP) dispatches, we are currently developing an ontological resource[11)]. This resource is designed to collect, to organize and to exploit the information extracted from dispatches for purposes of categorization, reference retrieval or thematic filtering, etc. Named entities are the main data of this resource. The population is progressively carried out by the module called "Resolution"[12)], using NEs previously extracted from AFP corpora.

Fig. 1 presents the framework of the ontology enrichment procedure.

To populate the ontological resource, we first worked on relation extraction from texts. For the automatic acquisition of relations and relation patterns, we implemented a clustering-based unsupervised method[5)].

We use AFP parsed corpora for the extraction of relations and relation patterns. The parsing (see 1 in Fig. 1) is carried out by the FRMG parser[13)], which uses a named entity recognition module, SxPipe[9)], to identify and classify NEs. From the parsing result, our RE system (see 2 in Fig. 1) builds syntactic dependency

trees and identifies all NE pairs and the paths we call "relation syntactic paths" that connect these NE pairs. We assume that the paths connecting NE pairs represent their relation. In the sentence

"*Xavier Bertrand succède officiellement Patrick Devedjian*"
(*Xavier Bertrand formally succeeds Patrick Devedjian*)

we identify an NE pair (*Xavier Bertrand*, *Patrick Devedjian*) and the path, $(X) \rightarrow^{SUJ-V} succder_v \leftarrow^{CPL-V(\text{à})} (Y)$, which connects this pair and represents its relation.

The system provides two types of results: on the one hand a list of classes containing relation paths (see 4 in Fig. 1) and on the other a list of classes containing NE pairs connected by a path which belongs to one of the path classes (see 3 in Fig. 1). The classes correspond to the relations and the paths correspond to patterns of the relation the class represents. The NE pairs correspond to instances of the relation represented by the class to which the pair belongs.

The integration of acquired relation instances into the ontology is already in progress[6]. In the next section, we show how the extracted patterns are exploited for the new named entity acquisition.

3. Named entity acquisition using relation patterns

We propose two NE acquisition methods using relation patterns. The first uses relation syntactic paths to identify NEs in a parsed corpus (see 5 in Fig. 1). The second method (see 6 in Fig. 1) consists in transforming the relation syntactic paths into string patterns in order to identify NEs in unparsed texts.

3.1 Using syntactic paths as extraction patterns

The first approach exploits, as in the work of Snow et al.[10], relation syntactic paths to locate new NEs by pattern matching. The operation consists of

(1) constructing syntactic trees from a parsed corpus,
(2) identifying the paths corresponding to an input pattern, and
(3) acquiring the items located at both ends of the paths.

When we identify the elements connected by a path corresponding to one of the input patterns, we assume that on the one hand they are NEs and on the other the type of these entities corresponds to one defined in the pattern. Consider, for example, a path representing the "CEO" relation that links two NEs, person type one X (NE_IND) and company type one Y (NE_COM):

$$X(NE\_IND) \rightarrow^{APPOS} PDG \leftarrow^{MOD-N(de)} Y(NE\_COM)$$

In the parse tree representing the following sentence, the syntactic structure of the underlined strings corresponds to this CEO pattern.

M. Rousset, qui reconnaît avoir "une divergence d'appréciation" sur le dossier Toyal avec M.Lassalle, doit rencontrer le PDG de Toyo, Masao Imasu, vendredi à Osaka (ouest du Japon) pour le "rassurer".

The pattern matching result is *X = Masao Imasu (pn)* and *Y = Toyo (pn)*. As the types of the elements connected by the path do not correspond to the types defined in the pattern, the system proposes a reclassification and provides two new NEs as a result: *X = Masao Imasu (NE_IND)* and *Y = Toyo (NE_COM)*.

The elements linked by the paths can be not only simple units but also compound units. In addition to pure simple proper names, there are also NEs composed of pure proper names and common nouns, or only common nouns, possibly with adjectives. They are sometimes called mixed proper names or descriptive-based proper names[3]. In order to identify these compound units, all dependent items are included in the identified NEs connected by a path. Consider the following sentence:

France Gamerre, 64 ans, présidente de Génération Ecologie.

By the matching it with a "President" pattern (person-organization relation), *France (location)* and *Génération (n)* are indentified as the two connected items. But after including dependent items, the result finally becomes *X = France Gamerre (NE_IND)* and *Y = Génération Ecologie (NE_ORG)*.

3.2 Transformation of syntactic paths into string patterns

The second approach consists of the transformation of syntactic paths into string patterns for acquiring new NEs during the dynamic procedure. String patterns allow the identification of NEs in unparsed corpora.

The transformation of syntactic paths into string patterns is carried out using the contexts in which these patterns occur. The operation consists in locating in context examples two NEs connected by the path and extracting the strings which contain all words included on the syntactic path. These strings can be located either

- between the two identified NEs or

- between a word that consistutes a node of the syntactic path used for matching and one of two identified NEs.

Take, for example, a path of the "Director" relation that links two NEs, person type one X (NE_IND) and organization type one Y (NE_ORG):

$X(NE\_IND) \rightarrow^{APPOS}$ *directeur* $\leftarrow^{MOD-N(de)} Y(NE\_ORG)$.

The path links instances of this relation in many different contexts. Consider the following texts:

1. *Christian Larièpe, directeur technique du FC Nantes, écope d'un match d'interdiction de banc de touche et de vestiaire d'arbitres pour son comportement envers l'arbitre.*
2. *Il s'agit de Christian Charpy, actuel directeur général de l'ANPE, ex-directeur de cabinet de Philippe Douste-Blazy.*

Two NEs, *Christian Larièpe* and *Nantes*, are connected by the path in text 1 and *Christian Charpy* and *ANPE* in text 2. Extracting the strings between the two linked NEs, we can derive the following string patterns (1b from text 1, 2b from text 2) for this syntactic path:

1b. *[EN_IND] , directeur technique de [EN_ORG]*
2b. *[EN_IND] , actuel directeur général de [EN_ORG]*

In the same way, another path of the "Director" relation

$X(NE\_IND) \leftarrow^{APPOS}$ *directeur* $\leftarrow^{MOD-N(de)} Y(NE\_ORG)$

also occurs in many different contexts and links instances in the following contexts:

3. *Le directeur des recherches, Carles Lalueza Fox, de l'Université de Barcelone, émet l'hypothèse qu'ils avaient la possibilité de détecter d'autres composants pas encore identifiés, ce qui permettrait d'expliquer la raison de la survivance du gène récessif.*
4. *Le directeur départemental de La Poste, Jean-Claude Sénat, a toutefois évoqué auprès de l'AFP une "première étape d'organisation dès 2009".*

From these examples, the system extracts strings between the word "directeur" and the NEs which occur in the right context of the other, so that all words located on the syntactic path are included (3b from text 3, 4b from text 4):

3b. *directeur des recherches , [EN_IND] , de [EN_ORG]*

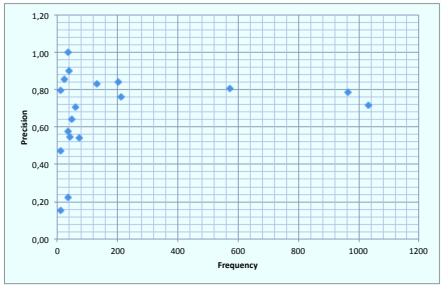

Fig. 2 Relationship between frequency and precision

4b. *directeur départemental de [EN_ORG] , [EN_IND]*

4. Experiment

4.1 Syntactic path method

An experiment was carried out to examine the performance of the syntactic path method with 18 patterns (see A.1) and AFP corpus consisting of dispatches from January 2007. These patterns are extracted by our RE method from the corpus of AFP dispatches from 2009. The corpus used for the experiment contains 565 291 sentences and our program found 4 077 paths matching a pattern.

The result of the experiment is presented in **Table 1**. Each column corresponds to the result obtained with each input pattern. The frequency indicates the number of extracted paths that provide new NEs and the precision is defined as the proportion of the correct results against all the extracted new ENs.

As can be seen in the result (Table 1) and in **Fig. 2**, high-frequency patterns provide interesting results, and not all low-frequency patterns are necessarily inefficient. Some low-frequency patterns also provide very interesting results. To identify effective patterns is a delicate task and selection criteria remain to be defined.

This approach using syntactic paths has an advantage of extracting distant NE pairs in the linear representation of texts which are difficult to identify. Another advantage of this approach is to make it possible to identify compound NEs (see **Fig. 3**). Compound NEs are not often contained in the lexical resources, and

Table 1 Extraction precision of each pattern

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Frequency	1033	965	131	203	572	213	61	25	35	38	47	43	72	13	13	35	37	13
Precision	0.72	0.79	0.83	0.84	0.80	0.76	0.71	0.85	1.00	0.90	0.64	0.55	0.54	0.15	0.47	0.58	0.22	0.79

Haut comité pour le logement des personnes défavorisées
Fédération Bosniaque de football
Fédération française d'athlétisme
Caisse nationale d'assurance maladie
Mouvement de libération du Soudan (SLM)
Syndicat des médecins libéraux (SML)
l'Union Nationale pour la Prévention du Suicide (UNPS)
l'Union européenne de football (UEFA)

Fig. 3 Examples of extracted compound NEs

it is difficult to identify these units by an annotation method based on lexical resources.

Most errors come from incorrect parsing results. Another source of error is the ambiguity of the element types for certain relationships. Relations such as "*Patron* (boss)" or "*Entraîneur* (coach)" involve not only person-organization pairs (*person X is head of organization Y*) but also person-person pairs (*person X is boss of person Y*). In most cases, the relations "*Directeur* (director)" and "*Président*" refer to person-organization pairs, but sometimes they involve person-event pairs (ex. *Aldo Tassone, directeur artistique du festival du film français de Florence*).

In addition, some extracted elements are anaphoric and require additional processing of anaphora resolution before these extracted NEs can be integrated into the resource.

4.2 Transformation-based approach

In order to examine what we could obtain by transformation, we transformed syntactic paths that belong to nine classes obtained by our RE method. **Table 2** shows the result of this experiment. The "Total obtained" row indicates the total number of obtained string patterns. The "NE tag error" and "Parse error" rows contain the number of incorrect patterns, due respectively to NE annotation error and parse error. The last row, "Patterns too specific", indicates the number of patterns which contain phrases too specific, especially with parenthetical clauses or relative clauses.

[NE_IND] , PDG de [NE_COM]
[NE_IND] , PDG du groupe [NE_COM]
PDG de [NE_COM] , [NE_IND]
[NE_IND] , le dernier PDG de [NE_COM]
[NE_IND] , l'actuel PDG de [NE_COM]
[NE_IND] , actuel PDG de [NE_COM]
[NE_IND] , ancien PDG de [NE_COM]
[NE_IND] , nouveau PDG [NE_COM]
[NE_IND] , le PDG de [NE_COM]
[NE_IND] , le nouveau PDG de la [NE_COM]
[NE_IND] , un ex-PDG de [NE_COM]
[NE_IND] , ex-PDG de [NE_COM]
[NE_IND] (PDG de [NE_COM]
[NE_IND] (ex-PDG de [NE_COM]
[NE_IND] (le PDG de [NE_COM]
[NE_IND] , PDG du groupe français [NE_COM]
[NE_IND] , le PDG du groupe français [NE_COM]
[NE_IND] , alors futur PDG de [NE_COM]
[NE_IND] , qui a par ailleurs été PDG du [NE_COM]
[NE_IND] , qui est aussi PDG de [NE_COM]
PDG de [NE_COM] à l'époque , [NE_IND]

Fig. 4 Examples of obtained string patterns: Class PDG

Except for the class "Président" which contains many paths created with parsing errors, we obtained a large number of correct and interesting patterns (see **Fig. 4**, **Fig. 5**).

These string patterns could make it possible to identify unknown NEs in unparsed texts and avoid errors due to incorrect parsing results. However, our work has not yet been completed. We have to define a grammar to integrate the unknown NEs identification into the dynamic procedure for AFP dispatch processing and then evaluate the performance of our string pattern-based identification method.

5. Conclusion

We have yet traced in outline our automatic named entity acquisition methods.

Table 2 Result of pattern transformation

Class	Président (president)	Directeur (director)	Directrice (director)	Entraîneur (coach)	Économiste (economiste)	Professeur (profesor)	PDG (CEO)	Patron (boss)
Total obtained	111	170	59	102	24	56	81	90
NE tag error	0.05	0.09	0.05	0.04	0.04	0.02	0.05	0.09
Parse error	0.35	0.04	0	0.04	0.13	0.05	0.07	0
Patterns too specific	0.27	0.11	0.32	0.13	0.08	0.04	0.15	0.12

[NE_IND] , professeur de Sciences Politiques à [NE_ORG]
[NE_IND] , professeur d'économie à l'université [NE_ORG]
[NE_IND] , professeur d'ingéniérie biochimique à [NE_ORG]
[NE_IND] , professeur en relations internationales et spécialiste de [NE_ORG]
[NE_IND] , professeur de philosophie politique [NE_ORG]
[NE_IND] , professeur d'archéologie préhistorique à [NE_ORG]
[NE_IND] , professeur de droit financier à [NE_ORG]
[NE_IND] , professeur de science environnementale à [NE_ORG]
[NE_IND] , professeur de droit constitutionnel à [NE_ORG]
[NE_IND] , professeur de droit international à [NE_ORG]
[NE_IND] , professeur de droit fiscal à [NE_ORG]
[NE_IND] , professeur d'histoire religieuse amricaine à [NE_ORG]
[NE_IND] , professeur de sociologie militaire à [NE_ORG]
[NE_IND] , professeur en ocanographie gologique à [NE_ORG]
[NE_IND] , professeur de microbiologie clinique à [NE_ORG]
professeur [NE_IND] de [NE_ORG]
professeur d'anthropologie [NE_IND] , de [NE_ORG]
professeur [NE_IND] ([NE_ORG]

Fig. 5 Examples of obtained string patterns: Class Professeur

This whole process is twofold. The exploitation of syntactic paths as extraction patterns allows the discovery of distant NE pairs which are difficult to identify in the linear representation of texts. But the use of parsing results is always accompanied by risks due to parsing errors. However, the transformation strategy produces string patterns, which allow us to avoid the influence of parse quality, giving robustness against the inconveniences of parsing to our model.

Acknowledgment

This work is funded by and conducted under the SCRIBO project and the EDyLex project.

References

1) Collins, M. and Singer Y.: Unsupervised Models for Named Entity Classification, *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.100–110 (1999).
2) Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Weld, D. S. and Yates, A.: Unsupervised Named-Entity Extraction from the Web: An Experimental Study, *Artificial Intelligence*, 165(1), pp.91–134 (2005).
3) Jonasson, Kerstin.: *Le nom propre. Constructions et interprtations*, Louvain-la-Neuve, Duculot (1994).
4) Komachi, M. and Suzuki, K.: Improving Semi-supervised Acquisition of Semantic Knowledge from Query Logs, *Transactions of the Japanese Society for Artificial Intelligence*, 23(3), pp.217–225, in japanese (2008)
5) Nakamura-Delloye, Y.: Extraction non-supervisée de relations basée sur la dualité de la représentation, *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montpellier, France (2011).
6) Nakamura-Delloye, Y. and Stern, R. Extraction de relations et de patrons de relations entre entités nommées en vue de lenrichissement dune ontologie, *Actes de TOTh 2011 (Terminologie & Ontologie : Théories et applications)*, Annecy, France (2011).
7) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations, *COLING/ACL-06*, pp.113–120 (2006).
8) Pennacchiotti, M. and Pantel, P.: Entity extraction via ensemble semantics, *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-09)*, Singapore, pp.238–247 (2009).
9) Sagot, B. and Boullier, P.: SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts, *Traitement Automatique des Langues (T.A.L.)*, 49(2), pp.155–188 (2008).
10) Snow, R., Jurafsky, D. and Ng, Y. A.: Learning syntactic patterns for automatic hypernym discovery, *Advances in Neural Information Processing Systems (NIPS 2004)* (2004).
11) Stern, R. and Sagot, B.: Resources for named entity recognition and resolution in

news wires, *Proceedings of LREC 2010 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, La Valette, Malte (2010).

12) Stern, R. and Sagot, B.: Détection et résolution d'entités nommées dans des dépêches d'agence, *Actes de TALN 2010(Traitement automatique des langues naturelles)*, Montréal, Canada (2010).

13) Villemonte de la Clergerie, E., Sagot, B., Nicolas, L. and Guénot, M.-L.: FRMG : evolutions dun analyseur syntaxique tag du francais *Journée ATALA "Quels analyseurs syntaxiques pour le francais ?"* (2009).

Appendix

A.1 Syntactic patterns used for the experiment

- Pattern 1 : $(PERSON) \leftarrow^{APPOS} \textit{président}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 2 : $(PERSON) \rightarrow^{APPOS} \textit{président}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 3 : $(PERSON) \leftarrow^{APPOS} \textit{entraîneur}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 4 : $(PERSON) \rightarrow^{APPOS} \textit{entraîneur}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 5 : $(PERSON) \rightarrow^{APPOS} \textit{directeur}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 6 : $(PERSON) \leftarrow^{APPOS} \textit{directeur}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 7 : $(PERSON) \rightarrow^{APPOS} \textit{directrice}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 8 : $(PERSON) \rightarrow^{APPOS} \textit{analyste}_{nc} \leftarrow^{MOD-N(\text{à})} (ORGANIZATION)$
- Pattern 9 : $(PERSON) \leftarrow^{APPOS} \textit{analyste}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 10 : $(PERSON) \leftarrow^{APPOS} PDG_{nc} \leftarrow^{MOD-N(\text{de})} (COMPANY)$
- Pattern 11 : $(PERSON) \rightarrow^{APPOS} PDG_{nc} \leftarrow^{MOD-N(\text{de})} (COMPANY)$
- Pattern 12 : $(PERSON) \leftarrow^{APPOS} \textit{patron}_{nc} \leftarrow^{MOD-N(\text{de})} (COMPANY)$
- Pattern 13 : $(PERSON) \rightarrow^{APPOS} \textit{patron}_{nc} \leftarrow^{MOD-N(\text{de})} (COMPANY)$
- Pattern 14 : $(PERSON) \rightarrow^{JUXT} \textit{patron}_{nc} \leftarrow^{MOD-N(\text{de})} (COMPANY)$
- Pattern 15 : $(PERSON) \leftarrow^{APPOS} \textit{photographe}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 16 : $(PERSON) \rightarrow^{APPOS} \textit{économiste}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 17 : $(PERSON) \leftarrow^{APPOS} \textit{professeur}_{nc} \leftarrow^{MOD-N(\text{de})} (ORGANIZATION)$
- Pattern 18 : $(PERSON) \rightarrow^{APPOS} \textit{professeur}_{nc} \leftarrow^{MOD-N(\text{à})} (ORGANIZATION)$