Recherche de motifs de graphe en ligne

Bruno Guillaume
LORIA, Inria Nancy Grand-Est *
bruno.guillaume@loria.fr

Résumé. Nous présentons un outil en ligne de recherche de graphes dans des corpus annotés en syntaxe.

Abstract.
Online Graph Matching
We present an online tool for graph pattern matching in syntactically annotated corpora.

Mots-clés : Syntaxe de dépendances, Corpus, Graphes.

Keywords: Dependency Syntax, Corpus, Graph matching.

Contexte

Les annotations linguistiques, par exemple en syntaxe sont souvent représentées par des arbres, soit en constituants, soit en dépendances. Le fait de se retreindre aux arbres a des avantages pratiques notamment pour calculer ces structures. Cependant, du point de vue linguistique, les arbres ne sont souvent pas suffisants lorsque l'on veut enrichir les structures. Le corpus DEEP-SEQUOIA (Candito *et al.*, 2014), par exemple, propose une annotation en dépendances profondes de phrases en français. Dans ce corpus, aucune hypothèse n'est faite sur les structures employées et il y donc de très nombreux cas d'annotations qui ne se représentent pas comme des arbres : par exemple certaines unités lexicales ont plusieurs gouverneurs (jusqu'à 7 dans la version 1.1 du corpus) et il existe de nombreux cycles.

C'est pour ces raisons que nous avons proposé d'utiliser la réécriture de graphes comme cadre formel pour décrire des processus de transformations de structures syntaxiques. Le logiciel GREW (Guillaume *et al.*, 2012) implémente ce modèle de calcul et permet de faire ce type de transformation. Pour déclencher l'application d'une règle, GREW utilise une recherche de motifs de graphes (pattern matching). C'est cette fonctionnalité de GREW qui est exploitée dans la version en ligne GREW-WEB [1]. Dans cet outil, on écrit un motif de graphe (généralement un petit graphe) et on peut visualiser les occurrences correspondantes dans un corpus donné. GREW-WEB est disponible avec quelques corpus libres de droits : SEQUOIA (Candito & Seddah, 2012), DEEP-SEQUOIA (Candito & Seddah, 2012) en français, UNIVERSAL DEPENDENCY TREEBANK (McDonald *et al.*, 2013) en français et en coréen et TIGER (Brants *et al.*, 2004) en allemand.

Exemples de recherche

Recherche d'une sous-catégorisation On recherche, dans SEQUOIA, un verbe avec à la fois un argument `a_obj` et un argument `de_obj`. Le résultat obtenu (6 occurrences) est représenté dans la Figure 1.

```
match {  V [cat=V]; A []; DE [];                % les 3 nœuds recherchés
         V -[a_obj]-> A; V -[de_obj]-> DE; }    % les relations entre les nœuds
```

Utilisation des contraintes négatives On peut filtrer les résultats obtenus en ajoutant des contraintes négatives. Ici, on recherche, toujours dans SEQUOIA, les occurrences de *prendre* avec un objet nominal sans déterminant (11 occurrences).

```
match { V[lemma="prendre"]; OBJ[cat=N]; V -[obj]-> OBJ }  % "prendre" + OBJ nominal
without { D[]; N -[det]-> D }                             % sans det pour l'OBJ
```

*. Ce travail a bénéficié du soutien du projet Ortolang (`ortolang.fr`). L'auteur remercie Antoine Chemardin pour son aide dans le développement de l'interface Web.

1. `http://grew.loria.fr/demo`

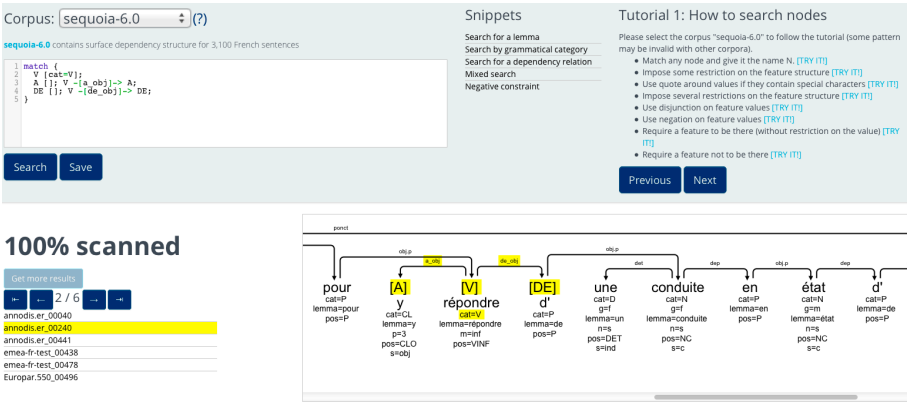

FIGURE 1 – Capture d'écran de l'interface

Recherche d'erreurs dans un corpus GREW-WEB permet de rechercher systématiquement des motifs qui sont susceptibles d'être des erreurs. Par exemple, dans SEQUOIA, on peut vérifier l'accord sujet-verbe. On trouve 23 occurrences du motif suivant, ce qui permet de repérer une dizaine d'erreurs d'annotation.

```
match { S [n=*]; V [cat=V, n=*]; V -[suj]-> S }                % le motif sujet-verbe
without {S.n = V.n}                                            % les traits "n" différents
without {V[m=part, t=past]; A[lemma=avoir]; V -[aux.tps]-> A}   % on élimine l'aux avoir
without {S[n=s]; V[n=p]; S -[coord]-> *}                       % pas de coord. comme sujet
without {S[cat=N, lemma="minorité"|"dizaine"|...] }            % exceptions lexicales
```

Recherche de graphes Dans DEEP-SEQUOIA, les structures sont des graphes et on peut donc rechercher des motifs qui sont eux-aussi des graphes. Ci-dessous, on recherche les cycles de longueur 8, on en trouve 2 occurrences dans le corpus.

```
match { N1[]; N2[]; N3[]; N4[]; N5[]; N6 []; N7 []; N8[];
N1 -> N2; N2 -> N3; N3 -> N4; N4 -> N5; N5 -> N6; N6 -> N7; N7 -> N8; N8 -> N1 }
```

Conclusion

L'outil en ligne GREW-WEB permet de trouver rapidement des exemples en corpus de constructions particulières ou de rechercher de façon systématique des erreurs d'annotation. En fait, rien ne restreint l'usage de GREW-WEB à des structures en dépendances, il peut être utilisé sur tout type de graphes comme des analyses en constituants, des graphes de représentation sémantique par exemple.

Références

BRANTS S., STEFANIE D., EISENBERG P., HANSEN S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). TIGER : Linguistic Interpretation of a German Corpus. *J. of Language and Computation*, **2**, 597–620.

CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & VILLEMONTE DE LA CLERGERIE É. (2014). Deep Syntax Annotation of the Sequoia French Treebank. In *LREC*, Reykjavik, Iceland.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN*, Grenoble, France.

GUILLAUME B., BONFANTE G., MASSON P., MOREY M. & PERRIER G. (2012). Grew : un outil de réécriture de graphes pour le TAL. In *12ième conférence TALN*, Grenoble, France : ATALA.

MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TACKSTROM O., BEDINI C., CASTELLO N. B. & LEE J. (2013). Universal dependency annotation for multilingual parsing. In *Proc. of ACL 2013*.