UNITEX/GRAMLAB: plateforme libre basée sur des lexiques et des grammaires pour le traitement des corpus textuels

Tita Kyriacopoulou*, Claude Martineau*, Cristian Martinez*

*5 boulevard Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2
{tita,claude.martineau,cristian.martinez}@univ-mlv.fr

Résumé. L'objectif de notre recherche est de répondre aux besoins croissants et divers d'extraction d'information pertinente exprimés par de nombreuses disciplines. Nous utilisons pour cela l'analyseur multilingue de corpus Unitex/GramLab développé à l'Université Paris-Est Marne-la-Vallée. Il fait appel à une approche symbolique et utilise des ressources linguistiques, dictionnaires électroniques et grammaires locales. Cette présentation ne constitue qu'une prise en main d'Unitex/GramLab et ne reflète que très partiellement les possibilités du logiciel et son champ d'utilisation, notamment pour l'extraction d'information, qui s'étend du monde de la recherche à celui de l'industrie.

1 Introduction

Un des objectifs de notre recherche est l'identification et l'extraction automatique d'information pertinente à partir de données textuelles provenant d'une multiplicité grandissante de domaines (littéraire, journalistique, scientifique, médical,technique, etc) et de sources (bases de données, bibliothèques numériques, blogues, etc). Afin de permettre un accès efficace à l'information et d'en simplifier l'utilisation nous devons prendre en compte le traitement de corpus de grande taille, la réduction du bruit contenu, le multilinguisme, ainsi que plusieurs tâches de traitement par l'ordinateur et l'utilisateur. C'est pourquoi il est nécessaire d'automatiser certains processus, notamment l'extraction d'entités nommées (noms propres, adresses, dates, etc). Cette analyse automatique est effectuée par des outils principalement issus de deux disciplines : l'informatique et le TAL (Traitement Automatique des Langues). Les deux disciplines fondent leurs analyses sur des techniques statistiques et/ou des connaissances et des ressources linguistiques.

UNITEX/GRAMLAB [1] utilise des ressources linguistiques même s'il doit évoluer vers un système hybride. Il est open source, multilingue [2], multiplateforme et permet d'analyser des textes en langue naturelle grâce à des ressources linguistiques telles que des dictionnaires électroniques et des grammaires locales. Ces dernières sont fondées sur la notion d'automate et de manière plus générale de réseau de transitions récursif (RTN) comportant des sorties. Ces grammaires sont représentées sous forme de graphes aisément réalisables grâce à un éditeur intégré.

1. UNITEX/GRAMLAB a été principalement développé par Sébastien Paumier (2001-2012). Son développement se poursuit grâce à une communauté de développeurs et de linguistes.

2. Français, anglais, . . ., grec, russe, arabe (écriture de droite à gauche), thaï et coréen (absence de séparateurs).

2 Ressources linguistiques

2.1 Les dictionnaires électroniques

Les dictionnaires électroniques utilisés par UNITEX/GRAMLAB obéissent au format (1). Nous donnons ci-dessous des exemples d'entrées simples comme composées dans lesquelles *Hum* et *Prof* indiquent qu'il s'agit de noms (*N*) humains et de profession.

(1) *Forme fléchie,forme canonique.catégorie gram.+infos. synt.-sém.+...+synt.-sém. :infos. flex.*

avocat,avocat.N+Hum+Prof:ms	avocat d'affaires,avocat d'affaires.N+Hum+Prof:ms
avocate,avocat.N+Hum+Prof:fs	avocate d'affaires,avocat d'affaires.N+Hum+Prof:fs
avocats,avocat.N+Hum+Prof:mp	avocats d'affaires,avocat d'affaires. N+Hum+Prof:mp
avocates,avocat.N+Hum+Prof:fp	avocates d'affaires,avocat d'affaires.N+Hum+Prof:fp

2.2 Les grammaires locales

Les grammaires locales permettent de décrire des patrons linguistiques à l'aide de graphes. Chaque chemin qui conduit de l'état initial à l'état final est un motif accepté. A titre d'exemple, la grammaire locale de la figure 1 permet reconnaître un certain nombre de patrons linguistiques. Nous donnons ci-après quelques phrases reconnues et non reconnues.

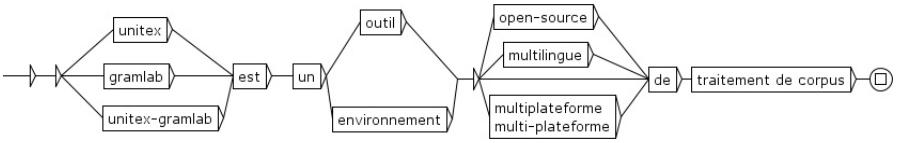


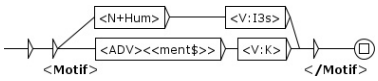
FIG. 1 – *Grammaire locale représentée sous forme de graphe*

— Unitex-GramLab est un outil multilingue de traitement corpus [**RECONNUE**]
— Unitex-GramLab est un outil de traitement de corpus [**RECONNUE**]
— Unitex-GramLab est difficile à apprendre [**ECHEC**]
— Unitex-GramLab est [**ECHEC**]

3 Recherche de motifs et annotation

3.1 Une grammaire qui fait appel au dictionnaire

Une grammaire peut faire appel au dictionnaire. Le graphe de la figure 2 comprend deux chemins, le premier reconnaît un nom humain <*N+Hum*> suivi d'un verbe à la troisième personne du singulier de l'imparfait <*V:I3s*>, le second reconnaît un adverbe qui se termine en *ment* <*ADV*><<*ment$*>> suivi d'un verbe au participe passé <*V:K*>. De plus cette grammaire a la capacité d'écrire dans le texte et entoure les séquences reconnues de balises *Motif* et de construire ainsi une annotation. Le résultat de l'application de cette grammaire se présente sous la forme d'une concordance dont un échantillon est visible à droite de la figure.

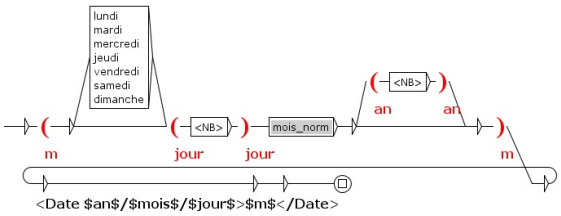


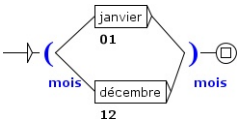
ecteur Fix, si <Motif>malencontreusement entraîné</Motif> dans un voya
'affaire. {S}L'<Motif>inspecteur de police était</Motif> donc extrêmem
ice était donc <Motif>extrêmement embarrassé</Motif>, quand la présenc

FIG. 2 – *Grammaire de reconnaissance de date et les concordances produites*

3.2 Utilisation d'un sous-graphe et de variables

La grammaire de la figure 3 effectue une normalisation d'une forme basique de date constituée au minimum du numéro de jour (le méta symbole *<NB>* reconnaît une suite de chiffres) et du mois en lettres, éventuellement précédé du nom de jour et/ou suivi de l'année. Elle utilise dans ce but les variables [3] *jour* et *an* qui mémorisent respectivement le numéro de jour et de l'année (si présente). La boîte grisée est un appel au sous graphe *mois_norm* qui transforme le nom du mois en son numéro [4]. La variable *m* mémorise la totalité de la date. Ensuite cette grammaire écrit une balise *<Date>* dans laquelle *an/$mois$/an* constitue la forme normalisée de la date (ordre année, mois, jour) et les *$* entourant chaque variable permettent d'afficher leur contenu. La figure 4 donne un extrait d'une concordance de dates normalisées et balisées.





Sous-graphe mois_norm (extrait)

FIG. 3 – *Grammaire de reconnaissance et de normalisation de dates*

Mr. Fogg, le <Date 1872/12/21>samedi 21 décembre 1872</Date>, à huit
cle parut le <Date /10/7>7 octobre</Date> dans le Bulletin de la Soc
Fogg. {S}Le <Date /10/9>mercredi 9 octobre</Date>, on attendait pou

FIG. 4 – *Concordance de dates normalisées*

3.3 Exemples d'annotations

La possibilité d'élaborer des grammaires utilisant de nombreux sous-graphes et niveaux de sous-graphes de sous-graphes confère à notre logiciel une grande puissance de description pour la construction de patrons linguistiques et d'extraction d'information. Nous pouvons citer certains travaux qui ont su en tirer parti par exemple pour l'extraction d'entités nommées (Maurel et al., 2011; Krstev et al., 2013), l'extraction de segments complexes qui étend la notion d'entités nommées (Kyriacopoulou et Martineau, 2015). La figure 5 donne un exemple d'annotation [5] d'entités nommées dans lequel les reconnaissances d'un nom de personne et d'une date sont mis en évidence.

3. Dans une grammaire la zone mémorisée dans une variable est délimitée par des parenthèses qui portent son nom en indice. Les parenthèses rouges représentent des variables d'entrées et les bleues de sorties.

4. Pour des raisons de place seul deux chemins sont exprimés.

5. Cette visualisation est obtenue à partir du texte balisé produit par UNITEX/GRAMLAB. Un script est utilisé pour transformer ce texte en un texte surligné de type html.

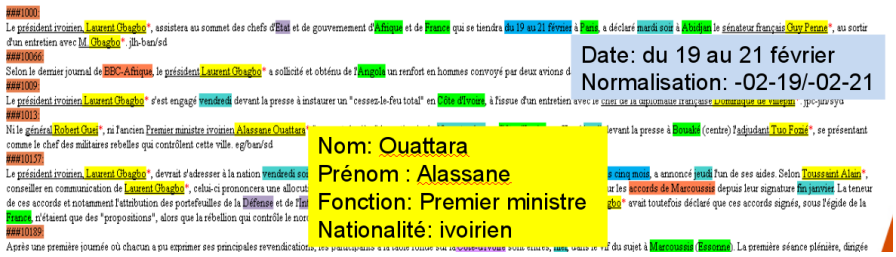


FIG. 5 – *Exemple d'annotation*

4 Conclusion

Nous avons effectué une présentation à la fois brève et détaillée de certaines possibilités d'Unitex/GramLab [6]. Notre logiciel s'utilise à travers deux interfaces écrits en Java UNITEX IDE (classique) et GRAMLAB IDE (orientée projet). Ils appellent le cœur du logiciel écrit en C/C++. Ce dernier est disponible sous la forme d'une API pour C et Java (JNI) qui de surcroît donne accès à un système de fichiers virtuels et à la persistance des ressources. Ces caractéristiques donnent la possibilité d'insérer d'UNITEX/GRAMLAB au sein de chaînes de traitement complexes. Il est utilisé par des universitaires (littéraires, linguistes, sociologues, etc) comme par des entreprises pour effectuer des tâches aussi diverses que l'analyse de textes littéraires, l'analyse de retours clients, la normalisation d'adresses, l'extraction d'opinions, etc. UNITEX/GRAMLAB n'est pas un logiciel dédié à une tâche particulière d'analyse de corpus ou d'extraction d'information mais grâce aux ressources dont il dispose et surtout grâce à celles dont il rend possible l'élaboration et l'utilisation, il permet à chacun de construire une solution la plus conforme à ses besoins et ses attentes.

Références

Krstev, C., I. Obradović, M. Utvić, et D. Vitas (2013). A system for named entity recognition based on local grammars. *Journal of Logic and Computation 24*(2), 473–489.

Kyriacopoulou, T. et C. Martineau (2015). Extraction de « segments complexes » : enrichissement des dictionnaires. *études de linguistique appliquée (éla) octobre-décembre 2015*(180), 407–416.

Maurel, D., N. Friburger, J.-Y. Antoine, I. Eshkol, et D. Nouvel (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues 52*(1), 69–96.

Summary

The goal of our research is to meet the growing needs of extraction of relevant information being faced by many disciplines. We present here a brief demonstration of UNITEX/GRAMLAB a software developed at the University of Paris-Est Marne-la-Vallée. It is a multilingual corpus analyzer which is based on a symbolic approach and uses linguistic resources, electronic dictionaries and local grammars. We only focus on some of the main features of UNITEX/GRAMLAB and we do not exhaustively explore their fields of application, especially to develop information extraction systems running over large-scale corpora.

6. Les binaires d'installation d'UNITEX/GRAMLAB sont disponibles sur `http://unitexgramlab.org` et les codes sources sur `https://github.com/UnitexGramLab`