Les boîtes à outils pour le TAL (GATE, NLTK, LinguaStream, Unitex, UIMA etc.)

Institut National des Langues et Civilisations Orientales

Contenu

Présentation des outils informatiques appelées “boîtes à outils”, utilisées pour la création de ressources linguistiques, de leur principe sous-jacent d'automates d'états finis déterministes, et des principales boites à outils disponibles sur le marché, avec une attention plus conséquente portée sur l'une d'elles: UNITEX.

Introduction: boîtes à outils et corpus

En TAL, les corpus ont pour applications :
⚠ la création de ressources linguistiques (e.g., lexiques)
⚠ la création d'outils (étiqueteurs, analyseurs, etc.) par apprentissage automatique
⚠ l'évaluation des outils existants

Pour être utile, un corpus doit pouvoir être “manipulé” de façon adéquate. Typiquement, on veut pouvoir :
⚠ segmenter : découper le texte en “segments” (mots)
⚠ normaliser (e.g., éliminer les majuscules)
⚠ faire des calculs statistiques sur les mots du corpus

I – les langages des boîtes à outils

La plupart des boîtes à outils sont écrites en langage JAVA ou C++, prinicpalement pour raison de pérennité.

Les langages de programmation utilisés (C/C++ et Java) sont choisis pour leur portabilité et leur stabilité d’utilisation dans le temps. Mais il existent des boîtes à outils qui sont écrites dans d'autres langages, comme Nooj, qui est écrit dans le langage de Microsoft .NET platform, la réponse de Microsoft à Java.

II – le LADL et Maurice Gross

Intéressé par les travaux de Chomsky et de Schützenberger, l'ingénieur Maurice Gross revient des Etats-Unis et publie *Notions sur les grammaires formelles* (1967). Ensuite il crée le Laboratoire d'Automatique Documentaire et Linguistique (LADL, laboratoire du CNRS).
Conscient que les études transformationnelles de Chomsky sont incomplètes pour une description exhaustive d'une langue, Maurice Gross va, au sein de son labo du LADL, entreprendre une classification de tous les verbes simples du français (pas moins de 5 000 verbes divisés en 15 000 emplois). Parallèlement, il dirigera des travaux semblables sur les noms et les adjectifs. Il encouragera des travaux similaires portant sur d'autres langues, au sein d'un réseau baptisé RELEX.

En conséquence, les travaux de Maurice Gross aboutiront à la création de dictionnaires électroniques.
L'approche de Gross privilégie une représentation par automates finis, qu'il appelle des "grammaires locales". Son approche aboutit à une vaste bibliothèque de graphes, chacun de ces graphes décrivant un ensemble de combinaisons d'éléments plus ou moins figés couvrant un domaine syntaxique ou

sémantique particulier. En outre, les automates d'états finis permettent d'unifier la formalisation de tout le système, grammaires et dictionnaires, et de donner une représentation claire des ambiguïtés.

Ce vaste ensemble de données (lexique-grammaire, dictionnaires, bibliothèque de graphes) est formellement combinable au sein de programmes de traitement automatique du langage. Ceux-ci permettent une exploration très fructueuse de grands corpus, tant pour la recherche linguistique que pour l'extraction d'informations, et peut-être pour un début de traduction automatique. C'est à l'élaboration de ce puissant outil de traitement du langage, combinant toutes les données accumulées sur plus de quarante ans, que M. Gross travaillait les derniers temps.

III - les automates d'états finis déterministes

Un automate fini (parfois qualifié de « machine à états finis ») est une machine abstraite utilisée en théorie de la calculabilité ainsi que dans l'étude des langages formels. C'est un outil essentiel en informatique, par exemple son rôle dans la compilation des langages informatiques (procédé permettant de passer d'un langage de haut niveau en langage binaire, dit langage machine).

Un automate est constitué d'états et de transitions. Son comportement est dirigé par un mot fourni en entrée : l'automate passe d'état en état, suivant les transitions, à la lecture de chaque lettre de l'entrée. L'automate est dit « fini » car il possède un nombre fini d'états distincts, ne disposant donc que d'une mémoire bornée.

Contrairement aux méthodes statistiques, la constitution d'automates ne nécessite pas de corpus. Seule l'intervention d'un expert est nécessaire pour la constitution de règles linguistiques permettant la désambiguïsation d'une séquence.

L'étiqueteur du projet OuRAL intègre un module de désambiguïsation lexicale à base d'automates. Ces automates peuvent être édités avec l'outil Unitex.

Les automates permettent de définir des séquences d'opérations booléennes sur les catégories lexicales, formes ou lemmes. Ainsi lorsqu'une séquence est détectée (un chemin possible dans l'automate), l'analyse est automatiquement désambiguïsée.

IV – INTEX

INTEX est un environnement de développement linguistique permettant la construction, le test et la gestion des descriptions formalisées à large couverture des langues naturelles, sous forme de dictionnaires et de grammaires électroniques. INTEX se trouve donc être un outil de base pour les applications du TAL. Il existe des modules linguistiques *INTEX* pour une douzaine de langues, et une demi-douzaine d'applications informatiques du TAL ont été construites avec *INTEX* pour des exemples d'utilisation du logiciel.

Les fonctionnalités d'INTEX étaient particulièrement bien adaptées à un public de linguistes (description de la morphologie et de la syntaxe des langues), de documentalistes (analyse de corpus) et d'informaticiens du TAL (applications d'extraction d'information). Mais le logiciel a attiré très tôt des enseignants tant en linguistique qu'en FLE, ses principales qualités pour une utilisation pédagogique étant les suivantes.

Son ouverture. On ne trouve ni traitement probabiliste, ni "boite noire" dans INTEX. Ses fonctionnalités sont toutes guidées par les notions d'accessibilité et de paramétrabilité. En outre, les résultats de traitement, y compris intermédiaires, sont immédiatement compréhensibles.

Sa grande facilité d'utilisation. Il est aisé de construire sa première grammaire locale en dix minutes. Un enseignant non informaticien peut adapter rapidement les dictionnaires et les

grammaires du système pour les appliquer, avec les apprenants, à des corpus de textes de son choix.
La compatibilité et la disponibilité de données linguistiques à large couverture pour une douzaine de langues.

V – les types de données incluses dans les boîtes à outils (e.g. dans UNITEX)

Unitex est basé sur des données linguistiques précises et exhaustives. Trois types de données:
- Grammaires locales (ou graphes)
- Dictionnaires électroniques de mots simples
  (DELAF) et de mots composés (DELACF)
- Tables de lexique-grammaire (à terme)

Définitions
● Mot simple : une séquence de lettres; délimitation par des séparateurs (espaces, ponctuation, etc.)
● Mot composé : une séquence de mots simples, dont le sens est non compositionnel (par ex. *cordon bleu, pomme de terre, belle famille, porte-manteau*)

Un dictionnaire est un ensemble d'entrées lexicales. Exemple d'une entrée lexicale:
– forme fléchie: *institutrice*
– forme de base (ou canonique): *instituteur*
– Catégorie grammaticale: nom (N)
– Informations flexionnelles (genre,nombre): fs (pour « féminin singulier »)
– traits sémantiques : Humain
● Exemple: institutrice,instituteur.N+Hum:fs

Construction des dictionnaires
● Construction d'un dictionnaire de formes canoniques (ou formes de base)
● Construction de modules de flexion automatique (transducteurs)
● A chaque forme de base, on associe une classe flexionnelle (un ensemble de règles)
● DELAS => Flexion automatique => DELAF

Grammaires locales
● Descriptions de formes linguistiques sous la forme de graphes (factorisation des formes)
● Quelles descriptions ?
– Extension des dictionnaires de mots composés
– Regroupement de formes complexes en classes sémantiques (ex. les dates)
– Descriptions de formes complexes, semi-figées à figées avec différentes variantes

VI – les principales boîtes à outils

a – GATE

Cette boîte à outils résulte d'un projet britannique. Elle est en cours de réécriture en JAVA. La version 2 de GATE (*General Architecture for Text Engineering*) est sortie le 14 mars 2002. GATE est une infrastructure de développement de traitement du langage humain. GATE est développé par l'Université de Sheffield depuis 1995 et est utilisé dans une vaste variété de recherche et de projets de développements, incluant l'extraction d'information en anglais, bulgare, roumain, bengali, grec, espagnol, suédois, allemand, italien et français.
Le but de GATE est l'aide aux scientifiques et programmeurs dans trois directions:
- en spécifiant une architecture ou structure organisationnelle orientée vers le traitement du langage.
- en fournissant une framework qui puisse implémenter l'architecture et qui puisse être utilisés pour l'exploitation des traitements linguistiques dans diverses applications.
- en fournissant un environnement de développement.

b – UIMA

UIMA (Unstructured Information Management Architecture) est le framework de traitement des données non structurées lancé par IBM. L'objectif de ce framework est la description des étapes de traitement d'un

document non structuré (texte, image, vidéo, etc.), en vue d'en extraire de façon automatique des informations structurées. Par contre, UIMA ne décrit ni la façon dont ces informations doivent être extraites du texte, ni la façon de s'en servir.

Si UIMA peut séduire par son architecture et sa prise en compte de nombreuses problématiques de façon native (réutilisation de composants, montée en charge et déploiement distribué, prise en compte des erreurs, etc.), il reste encore purement orienté vers les problématiques du text-mining (fouilles de textes) pur, et n'a pas (encore) amorcé le virage des ontologies et des métadonnées contrôlées; le fossé entre l'information (brute) extraite du texte et une ontologie ou une base RDF à alimenter restant à combler.

c – Linguastream

Développée au GREYC depuis 2001, LinguaStream est une plate-forme générique pour le TALN (traitement automatique des langues naturelles), fondée sur l'enrichissement incrémental des documents électroniques.

LinguaStream permet la conception et l'évaluation de chaînes de traitement complexes, par assemblage de modules d'analyse de types et de niveaux variés (morphologique, syntaxique, sémantique, discursif ou encore statistique).

De cette façon, chaque palier de la chaîne de traitement se traduit par la découverte et le marquage de nouvelles informations, sur lesquelles pourront s'appuyer les analyseurs subséquents. En fin de chaîne, différents outils permettent une visualisation des documents analysés et leurs annotations. La plate-forme propose différents mécanismes d'élaboration des composants de traitement (règles morphologiques, transducteurs, règles de production, grammaires d'unification, lexiques sémantiques, etc.). La plupart d'entre eux s'appuient sur des formalismes déclaratifs, certains étant fréquemment utilisés en TAL.

Chaque composant d'analyse est immédiatement réutilisable dans d'autres chaînes de traitement, et peut ainsi être remplacé par un autre composant fonctionnellement équivalent.

Une interface graphique se charge des différents aspects de l'élaboration d'une chaîne de traitement complète.

De plus, grâce à une API Java et au recours systématique des normes et outils XML, LinguaStream ainsi que les traitements élaborés avec son aide sont aisément extensibles et intégrables.

d – NooJ

a été créé par Max Silberztein, qui est resté plus de 10 ans au sein de UNITEX (INTEX). NooJ supporte plus de 100 formats.

VII - UNITEX

Fonctionnalités générales

Cette plate-forme permet de construire des ressources linguistiques comme par exemple des dictionnaires électroniques ou des grammaires, et de les utiliser pour effectuer des recherches complexes dans un corpus et de construire des concordances.

Autres fonctionnalités :

- Traitement de l'ambiguïté lexicale par automates d'états finis déterministes.
- Utilisation de tables de lexique-grammaire.

Le dictionnaire inclus dans UNITEX contient plus de 90 000 mots simples sous leur forme canonique, à laquelle sont associés diverses informations linguistiques, comme par exemple des codes flexionnels. La flexion du DELAS s'opère selon plus de 350 paradigmes différents, dont 150 verbaux, pour constituer le DELAF.

Pour les mots composés on retrouve le DELAC. Ce dictionnaire contient plus de 100 000 mots composés (90 000 noms, 15 000 constructions « être Prép N », 8000 adverbes, 500 conjonctions).

Contexte d'utilisation

Unitex est utilisé au laboratoire IGM-LabInfo qui s'occupe de la construction et de la maintenance de ressources linguistiques (dictionnaires et grammaires).

UNITEX permet également une exploitation de ces ressources en les appliquant sur des textes, ce qui permet la recherche d'expressions complexes et la construction de concordanciers.

Cet aspect du système se trouve à la base de nombreuses applications, parmi lesquelles on peut citer les plus importantes tels le repérage de séquences (par exemple, entités nommées) ou encore l'extraction d'informations, le filtrage ou le routage de documents, etc.

Références bibliographiques

http://alsic.revues.org/index336.html

http://www.limsi.fr/Individu/habert/Projets/JourneeATALA05ArticulerLesTraitements/linguastream/paper/final.pdf **[LinguaStream]**
http://infolingu.univ-mlv.fr/ **[UNITEX]**
http://gate.ac.uk/ **[GATE]**
http://www.nooj4nlp.net/ **[NooJ]**
http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html **[UIMA]**