Du caractère au sens : méthodes et outils pour le TAL

LDI

13 et 14 septembre 2011

Sommaire

Description

Le déroulement d'une application de traitement automatique des langues (TAL) des textes écrits peut être envisagé comme un processus qui se découpe en plusieurs étapes successives. Il s'agit de transformer une suite indistincte de caractères – lettres, espaces, ponctuations – en un ensemble cohérent d'éléments de nature linguistique – mots, mots composés, syntagmes, phrases... – où chacun de ces éléments se voit attachés un ou plusieurs traits – nature, sens, domaine... Ce travail une fois réalisé, il est alors possible d'effectuer différents traitements : traduction, extraction d'information... À chaque étape, l'application d'une ressource linguistique, quelle qu'en soit la nature, est nécessaire.

L'objectif de ces journées de formation organisées par le laboratoire LDI (Lexiques, Dictionnaires, Informatique) est de proposer un ensemble de modules abordant différents aspects liés à ce continuum. Chacun des cours est construit de façon à être le plus autonome possible. Ainsi, chaque participant peut choisir de construire son parcours « à la carte ». Afin de faciliter les choix des participants, chaque cours fera l'objet d'une description précise : les objectifs, les pré-requis nécessaires et les outils mis en œuvre.

Intervenants

- Lucie Barque
- Lou Burnard
- Pierre-André Buvet
- Emmanuel Cartier
- Fabrice Issac
- Sylvain Loiseau
- Céline Poudat
- Xavier-Laurent Salvador

Programme

mardi 13 septembre

- Construction et normalisation de corpus
- Etiquetage lexical
- Annotation sémantique de corpus
- Etiqueteurs morphos-syntaxiques

mercredi 14 septembre

- XML-TEI
- Corpus aligné, corpus comparables
- Etiqueteurs morphos-syntaxiques
- Analyse sémantique des textes
- Bases de données XML

Annotation sémantique de corpus

Formateur : Lucie Barque

Volume : 2h

L'annotation sémantique consiste à expliciter, au moyen d'un ensemble descriptif prédéfini, tout ou partie du contenu informationnel d'un texte. Le cours passera en revue les éléments du texte couramment annotés – étiquetage des entités nommées, annotation temporelle, apectuelle et modale des prédicats (cf. pour le français, les projets *French Time Bank* et *Nomage*), typage sémantique de leurs arguments (cf. pour l'anglais, *Propbank*), relations interphrastiques (cf. le projet COMTIS), etc – ainsi que les questions relatives aux différentes méthodes d'annotation (manuelle, semi-automatique ou automatique).

Liens

- http ://www.linguist.univ-paris-diderot.fr/ abittar/french-timebank/
- http ://sites.google.com/site/nomagesite/home
- http ://verbs.colorado.edu/ mpalmer/projects/ace.html
- http ://www.idiap.ch/project/comtis

Etiquetage lexical

Formateur : Pierre-André Buvet
Volume : 2h

Les applications dédiées au traitement de l'information textuelle intègrent des systèmes qui manipulent des textes ou des éléments de texte. Dans ces systèmes, on a longtemps privilégié les méthodes statistiques aux dépens des méthodes linguistique. Les premières ayant atteint un seuil qu'il paraît maintenant difficile de dépasser, les performances des systèmes qui les utilisent dépendent principalement de l'évolution technologique des machines. Les secondes ont comme principale particularité d'exploiter des ressources constituées dans la perspective du Traitement Automatique des Langues (TAL) : il s'agit principalement de banques de graphes, de bases de règles, de dictionnaires électroniques et d'ontologies. Dans le cadre de la formation, on présentera ces différentes ressources et on expliquera comment les utiliser pour faire de l'étiquetage lexical.

Outils

- UNITEX

Liens

- `http://www-igm.univ-mlv.fr/ unitex/`

Corpus aligné, corpus comparables

Formateur : Fabrice Issac
Volume : 3h

Les corpus multilingues sont des ressources indispensables dans de nombreuses applications liées au traitement automatique dans un contexte multilingue. La plus évidente de ces applications est la création de ressources terminologiques pour la traduction assistée ou automatique. Cependant, ce type de corpus permet aussi d'alimenter des logiciels de traduction automatique, des environnements d'apprentissage des langues (que ce soit comme support pour la génération d'exercices ou comme objet d'étude) et des outils de recherche d'information multilingue.

Une première définition d'un corpus multilingue est «un ensemble de textes dans plusieurs langues». Il est possible, à partir de cette définition simple, de distinguer différents catégories chacune liée plus particulièrement à un besoin spécifique. Ainsi un corpus multilingue parallèle ou bi-texte regroupe des textes en plusieurs langues . Un corpus multilingue aligné est un corpus multilingue parallèle où certains éléments textuels ou para-textuels – paragraphe, phrases, termes – sont mis en correspondance. Un corpus multilingue comparable est un corpus multilingue où le lien existant entre les textes dans des langues différentes est thématique et non plus une traduction directe.

Au cours de la formation nous présenterons un état de l'art lié à cette problématique et nous verrons concrètement comment construire et manipuler un corpus aligné.

Outils

- Jedit

Liens

- `http://www.revue-texto.net/Corpus/Publications/corpus-dh.pdf`

Base de Données xml

Formateur : Xavier-Laurent Salvador
Volume : 2h

Le traitement automatique de corpus nécessite la mise en place de dispositifs de description métalinguistiques fins et efficaces. XML répond parfaitement à tous les besoins en matière de traitement de corpus : la représentation des arborescences de données permet autant de structurer la représentation du corpus que les moyens d'accès à l'information. Mais le partage de l'information est d'autant plus difficile que la structure XML est conçue spécifiquement par un analyste. Les bases de données XML comme Basex, développé actuellement à l'Université de Konstanz par l'équipe de C. Grünn, sont un outil extrêmement efficace d'une remarquable simplicité, moyennant quelques connaissances acquises dans la construction de requête XQuery, et surtout très rapide pour reconfigurer un appareil textuel de plusieurs millions de lemmes en quelques secondes en fonction d'objectifs intégrés dans le cahier des charges de l'élaboration du corpus, mais aussi de répondre à des besoins très précis en termes d'accès à l'information.

Outils

- BaseX

Liens

- http ://basex.org

Analyse sémantique des textes

Formateur : Emmanuel Cartier
Volume : 3h

L'analyse sémantique est l'étape ultime de l'analyse linguistique des textes. Elle prend en compte les séquences de mots morpho-syntaxiquement annotés et tente de leur associer des significations. Nous évoquerons les différents algorithmes utilisés pour effectuer cette analyse sémantique, puis détaillerons les théories et méthodes basées sur des ressources linguistiques fines incluant des descriptions syntactico-sémantiques et des dictionnaires de relations sémantiques. Le cours évoquera les principaux problèmes à surmonter pour l'utilisation de ces ressources : prise en compte des phénomènes de mise en discours, gestion de la coréférence et de l'inférence. Des démonstrations accompagneront l'exposé théorique.

Outils

- WordNet
- FrameNet
- DicoValence
- TextBox

Liens

- http ://wordnet.princeton.edu/
- http ://bach.arts.kuleuven.be/dicovalence/
- http ://verbs.colorado.edu/verb-index/
- http ://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Presentation.html

Etiqueteurs morphosyntaxiques

Formateur : Céline Poudat
Volume : 3h

Le niveau morphosyntaxique joue un rôle crucial tant pour l'extraction de données ou la construction de ressources que pour la description linguistique des corpus. Il demeure l'un des niveaux linguistiques les plus développés et les plus automatisés, et de nombreux étiqueteurs morphosyntaxiques, ou taggers, sont disponibles. Nous dresserons un panorama comparé des outils disponibles (e.g. TreeTagger, TnT Tagger...), et inviterons les inscrits à cette formation à manipuler les outils abordés. La description d'un corpus ou la construction d'une ressource nécessitant de disposer d'observables linguistiques pertinents et adaptés à la tâche visée, nous aborderons finalement la question de l'entraînement des outils à un corpus et un jeu de descripteurs ad hoc.

Outils

–

Liens

- `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
- `http://www.coli.uni-saarland.de/ thorsten/tnt/`
- `http://ilk.uvt.nl/mbt/`
- `http://www.revue-texto.net/index.php?id=2293`