RLT: Acquisition de ressources linguistiques pour les TAGs

Eric de la Clergerie

Projet ATOLL – INRIA

28 Février 2000

Groupe de travail TAGML

Présentation

RLT est une Action de Recherche Concertée [ARC]

- sur 2 ans à partir de début 2001

- financée par l'INRIA (stagiaires DEA et postdoc, matériel et missions)

- Site WEB : `http://atoll.inria.fr/RLT`

- Liste de diffusion

Objectifs

Constitution semi-automatique du lexique syntaxique et du lexique morphologique pour la grammaire FTAG :

– lier les lemmes aux familles

– instancier les contraintes de traits et de co-ancrage

– lier les entrées morphologiques aux lemmes

Motivations

Dans le désordre :

- Mettre en valeur le travail de TAGML (dont les représentations XML).
- Recenser les outils et ressources pour les TAGs de nos différentes équipes et les faire coopérer.
- Améliorer l'analyse syntaxique des TAGs
- Se former sur les tâches d'acquisition semi-automatique
- Aider à la construction de lexiques pour le français
- Rechercher le bon niveau de description des lexiques (meta-grammaire)
- Aider à la diffusion de ressources linguistiques en facilitant leur création.

Participants

Quatre équipes sont impliquées :

Atoll (INRIA) – Eric de la Clergerie

Langue et Dialogue (LORIA) – Bertrand Gaiffe

Calligramme (LORIA) – Guy Perrier

TALANA (Paris 7) – Anne Abeillé

Atoll

Page Web : `http://www.inria.fr/Equipes/ATOLL-fra.html`

Membres de l'équipe impliqués :

- Éric de la Clergerie (Chargé de Recherche INRIA) – coordinateur
- Pierre Boullier (Directeur de Recherche INRIA)
- Philippe Deschamp (Chargé de Recherche INRIA)
- François Barthélemy (Maître de conférence au CNAM)

Contributions :

- Analyse syntaxique des TAGs
- Représentation XML de ressources linguistiques
  grammaires et forêts de dérivation
- Environnement TAG avec des outils XML
- Sensibilisation à la promotion de ressources « libres »

Langue et Dialogue

Page Web : `http://www.inria.fr/Equipes/LANGUEETDIALOGUE-fra.html`

Membres de l'équipe impliqués :

- Bertrand Gaiffe (Chargé de Recherche CNRS) – coordinateur
- Laurent Romary (Chargé de Recherche CNRS)
- Azim Roussanaly (Maître de conférences à l'université de Nancy 2)
- Samuel Cruz-Lara (Maître de conférences à l'université de Nancy 2)
- Jean-Luc Husson (Maître de conférences à l'université Henri Poincarré Nancy 1)
- Djamé Seddah (Doctorant)

Contributions :

- Gestion et distribution de corpus (SILFIDE, ELAN)
- Gestion de lexiques et dictionnaires (DHYDRO)
- Travail sur la méta-grammaire
- Outils TAGs (dont analyseurs)

Calligramme

Page Web : `http://www.inria.fr/Equipes/CALLIGRAMME-fra.html`

Membres de l'équipe impliqués :

- Guy Perrier (Maître de conférence à l'université de Nancy 2) – coordinateur
- Philippe de Groote (Chargé de Recherche INRIA)
- François Lamarche (Directeur de Recherche INRIA)

Contributions :

- Logique linéaire et analyse syntaxique
- Liens avec les meta-grammaires
- Formalisation des meta-grammaires

TALaNa

Page Web : `http://talana.linguist.jussieu.fr/`

Membres de l'équipe impliqués :

- Anne Abeillé (Professeur) – coordinateur
- Laura Kallmeyer (poste ATER)
- Kim Gerdes, Nicolas Barrier, Sébastien Barrier, Alexandra Kinyon, Lionel Clement (Doctorants)

Contributions :

- Expertise linguistique
- Grammaire et meta-grammaire du français
- Traitement des TAG – analyse, génération, « (super)tagging », ...
- Corpus annotés pour le français

Scénario

Grandes lignes :

1. Constitution de banques de dérivation ou corpus annotés suite à l'analyse syntaxique partielle de corpus avec des lexiques surgénératifs.
2. Utilisation de méthodes de classification (dont statistiques) pour faire émerger les emplois caractéristiques des mots.
3. Préparation de fiches « lexique » pour les entrées candidates
4. Soumission des fiches à un linguiste pour validation et correction au travers d'une interface (avec accès aux phrases analysées).
5. Boucle sur 1 ou 2 en prenant en compte les fiches validées jusqu'à obtention d'un point fixe.

Beaucoup de détails à examiner et à valider !

Programme de travail

– Constitution et mise en forme de corpus de textes
– Recensement des outils et ressources
– Mise en place des analyseurs syntaxiques (tests des meilleurs stratégies : analyse partielles ou complète, dirigée par les têtes, ...).
– Format des corpus annotés produits par les analyses
– Émergence des propriétés d'emploi des mots
– Rattachement de ces propriétés aux propriétés de la méta-grammaire
– Format des fiches lexiques (basées sur la DTD XML mais modifiées en relation avec la méta-grammaire).
– Fonctionnalités de l'interface de validation du linguiste.
– Stockage des lexiques et accès rapides, distribués (SILFIDE+serveur de grammaire).
– Mise en place de prototypes.
– Acquisition d'un petit lexique.

Outils et ressources nécessaires

Ressources :

- Corpus
- Grammaires ou meta-grammaire (schéma d'arbres)
- Lexique de base

Outils :

- Gestionnaire de corpus
- Taggers
- Analyseurs morphologique (mots inconnus + rattachement au lemme)
- Analyseurs syntaxiques
- Outils de classification
- Interface de constitution de lexique

Ouverture

RLT est un projet qui se veut ouvert :

- Réutilisation d'outils et ressources.
- Diffusion de ressources et d'outils
- Mise au point d'une méthodologie pour l'acquisition de ressources linguistiques transposable dans d'autres contextes.
- RLT point de départ pour d'autres projets plus ambitieux

$\Rightarrow$ Discussions et coopérations bienvenues avec d'autres intervenants.