Vers un scénario pour RLT

Eric de la Clergerie

Projet ATOLL – INRIA

INRIA

26 Septembre 2001

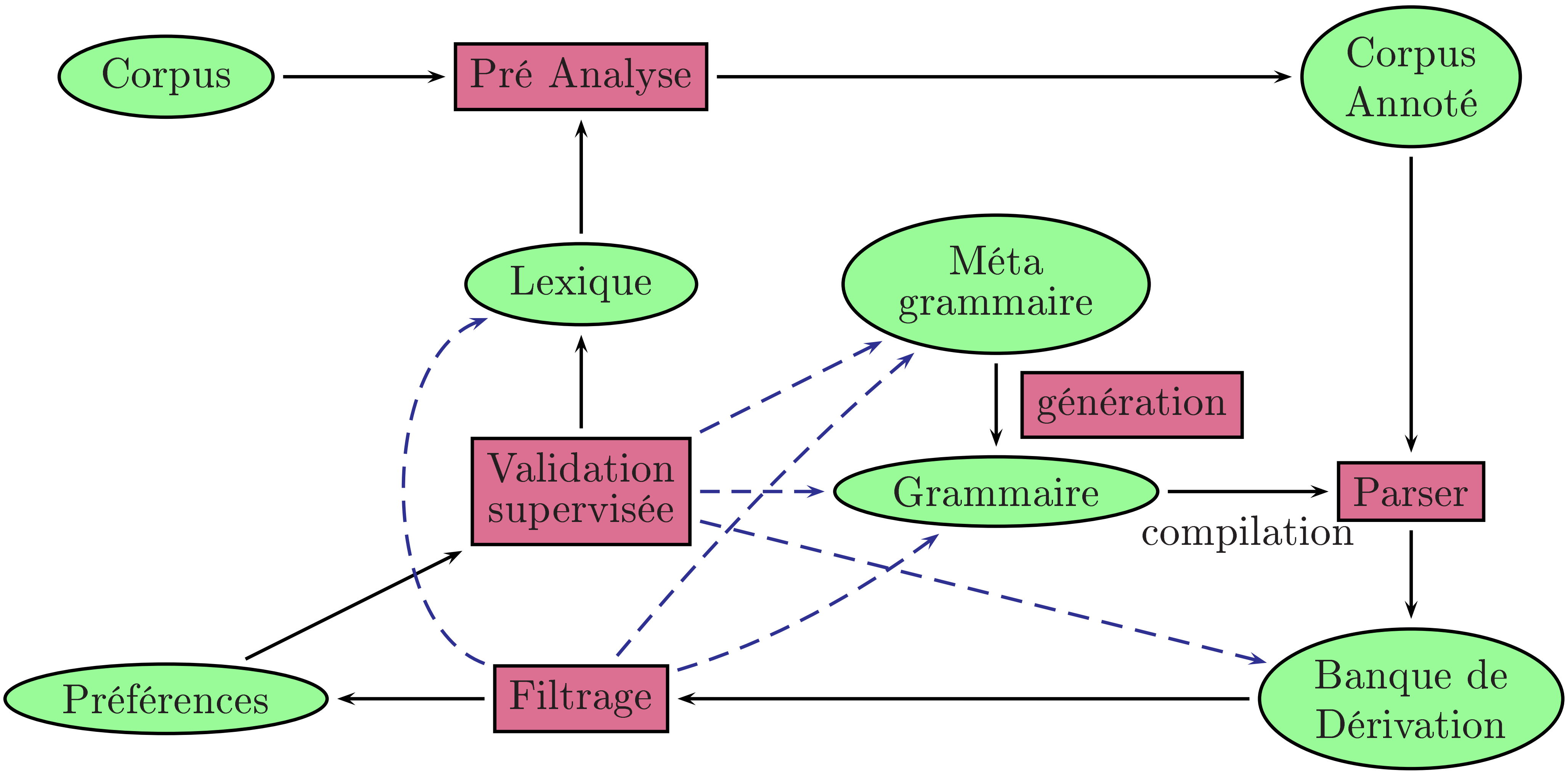
Réunion ARC RLT

Objectifs

1. Préciser le scénario d'acquisition à mettre en oeuvre pour l'action RLT

2. Support pour lancer la discussion

Architecture



Corpus

- Corpus du Monde
- Corpus INALF
- Corpus de phrases simples
- Corpus d'exemples pris dans un dictionnaire

Pre-traitement

Informations pouvant servir à guider les analyseurs :

Segmentation

Etiquetage Catégorie lexicale

Outils : Tree Tagger, Brill

Analyse Morphologique

Catégorie lexicale + racine + traits (genre, nombre, . . .)

Outils : Flemm

Chunking Frontières de constituants

Outils : Alexandra

Corpus annotés Structures des constituants (corpus TALaNa)

Corpus annoté : Format XML foundé sur celui de TALaNa

Grammaires

Plusieurs choix à envisager :

- Grammaire actuelle du Français de TALaNa
- Grammaire regénérée à partir de la méta-grammaire
  (directement générée en format XML)

Utilisation de :
- Grammaire complète
- Grammaire partielle (non prise en compte de phénomèmes rares)
- Grammaire jouet pour une première expérience

Parsing

– Parsing robuste (analyse partielle)

– Utilisation du parser RCG de Pierre Boulier
(mais éventuellement occasion de tests avec le parseur d'Anoop)
Pierre peut actuellement compiler des grammaires du plusieurs milliers d'arbres.

– Eventuellement, la sortie RCG guide un parseur DyALog
pour la prise en compte des traits et des co-ancrages.
DyALog compile de l'ordre de 1K arbres
(mais travail sur le compilateur en cours).

⇒ Banque de dérivation (en XML) gérées par le serveur de dérivation.

Filtrage

Objectif : Faire emerger les informations significatives, à partir des dérivations et de la grammaire (lexique + métagrammaire).

- arbres
- familles
- traits de la méta-grammaire
- instanciations
- co-ancrages

Des heuristiques à dégager !

Outils à développer en s'appuyant

1. sur le serveur de dérivation et son langage de requêtes.
2. sur le serveur de grammaire (avec les info de la méta-grammaire)

Filtrage (2)

Pour un lemme l donné :

- distribution pondérée des arbres utilisés
  - ratio arbre τ sur ensemble des arbres utilisés pour l
  - ratio derivations avec τ sur ensemble des dérivations l
  - prise en compte de la longueur des dérivations (en particuliers de celles qui sont totales)
  - prise en compte de mots du lexique dans les dérivations utilisées

- distribution pondérée des familles utilisées
  - ratio famille f sur ensemble des familles
  - distribution des arbres dans les familles couverture uniforme, couverture « piquée »

Filtrage (3)

Pour un lemme l donné (éventuellement un arbre ou une famille) :

- distribution des instanciations rencontrées
  instanciations similaires sur des noeuds ayant le même id (ou fonction).

- distribution des coancrages rencontrées
  mots ou lemmes rencontrés comme tête des noeuds de substitution.

Filtrage (4)

Alternative ou complément : co-specification

Identification de classe contextuelle de mots ayant

1. des distributions similaires d'arbres ou familles,
2. un représentant est dans le lexique
3. les têtes co-ancrés sont les mêmes (sur au moins une des fonctions à remplir)

⇒ attribution de la même famille aux mots de la classe

Validation

Définition d'une interface de validation :

– Grille de saisie pour entrer des mots dans le lexique
– Pré-remplie par les résultats du filtrage
– Pouvant être modifiée avant validation.
– La validation envoie ses résultats au serveur de grammaire (mode édition)

S'appuyer sur

– le travail de Benoît Crabbé (LORIA)
– les outils de visualisation d'arbres (analyse, dérivation, élémentaires)
  Eric de la Clergerie (ATOLL) et Christophe Cocquet (LORIA)

Validation (2)

Pour un lemme :

- liste ordonnée des familles avec info de distribution
- pour chaque famille,
  - distribution des arbres
  - affichage information niveau méta-grammaire
  - affichage informations d'instanciation et de coancrage
- pour une famille ou un arbre, affichage d'exemples,
  - exemple complet
  - exemple partiel
  - exemple simplifié (élimination adjonctions + génération)
  - exemples reconstruits pour des arbres non rencontrés
    (à partir des exemples rencontrés).

  Possibilité d'accéder aux forêts de dérivations et d'analyse (centrées sur le mot).
- possibilité de changer les paramètres proposés

Bootstrap

Après validation de nouvelles entrées dans le lexique,

1. remise à jour du corpus annoté et de la banque dérivation directement ou par réanalyse des phrases pertinentes

2. nouveau cycle de filtrage et validation

Formats

Corpus Texte, éventuellement balisé XML

Corpus annoté XML fondé sur le format du Corpus TALaNa (Lionel Clement)

Banque de dérivations XML, DTD Forêts + utilisation BD

Grammaire XML, DTD TAG2 + utilisation BD

Lexique XML, DTD TAG2 + utilisation BD

Métagrammaire A définir à partir du format actuel (migration XML ?)

Préférences A définir

Outils

Serveur de grammaires Compléter mode édition

Serveur de dérivations Compléter mode édition

Outils de visualisation Arbres élémentaires, d'analyse et de dérivation (ATOLL + LORIA)

Parsers Parsers RCG et DyALog.

Validation A partir du système de saisie du LORIA + outils de visualisation

Génération Grammaire Implantation Barrier ou Bertrand Gaiffe. Possibilité de générer la grammaire sous forme XML.

Filtreur A construire en exploitant serveurs de grammaires et de dérivation

Chunking Outils d'Alexandra

Tagging Tree Tagger, Bril, ...

Morphologie Flemm, ...