Algorithme de Earley pour les grammaires d'interaction

Jonathan Marchand

Mémoire de Master Informatique
spécialité Traitement Automatique des Langues

présenté et soutenu à Nancy, le 30 juin 2006

Introduction

- Le traitement automatique des langues (TAL) a pour objectif de traiter des données linguistiques exprimées dans une langue dite “naturelle”.
- Nous parlons ici plus précisément d'analyse syntaxique. Il s'agit d'expliciter la structure grammaticale de phrases sous forme d'arbres.
- Les grammaires d'interaction ont été conçues pour modéliser la syntaxe de la langue.
- Comment analyser un énoncé avec une grammaire d'interaction ?
- Idée : Reprendre uns stratégie d'analyse existante pour l'analyse de langages ambiguës comme la langue naturelle.
- L'algorithme de Earley est très efficace sur les grammaires hors-contextes et a fait ses preuves sur les grammaires d'arbres adjoints.

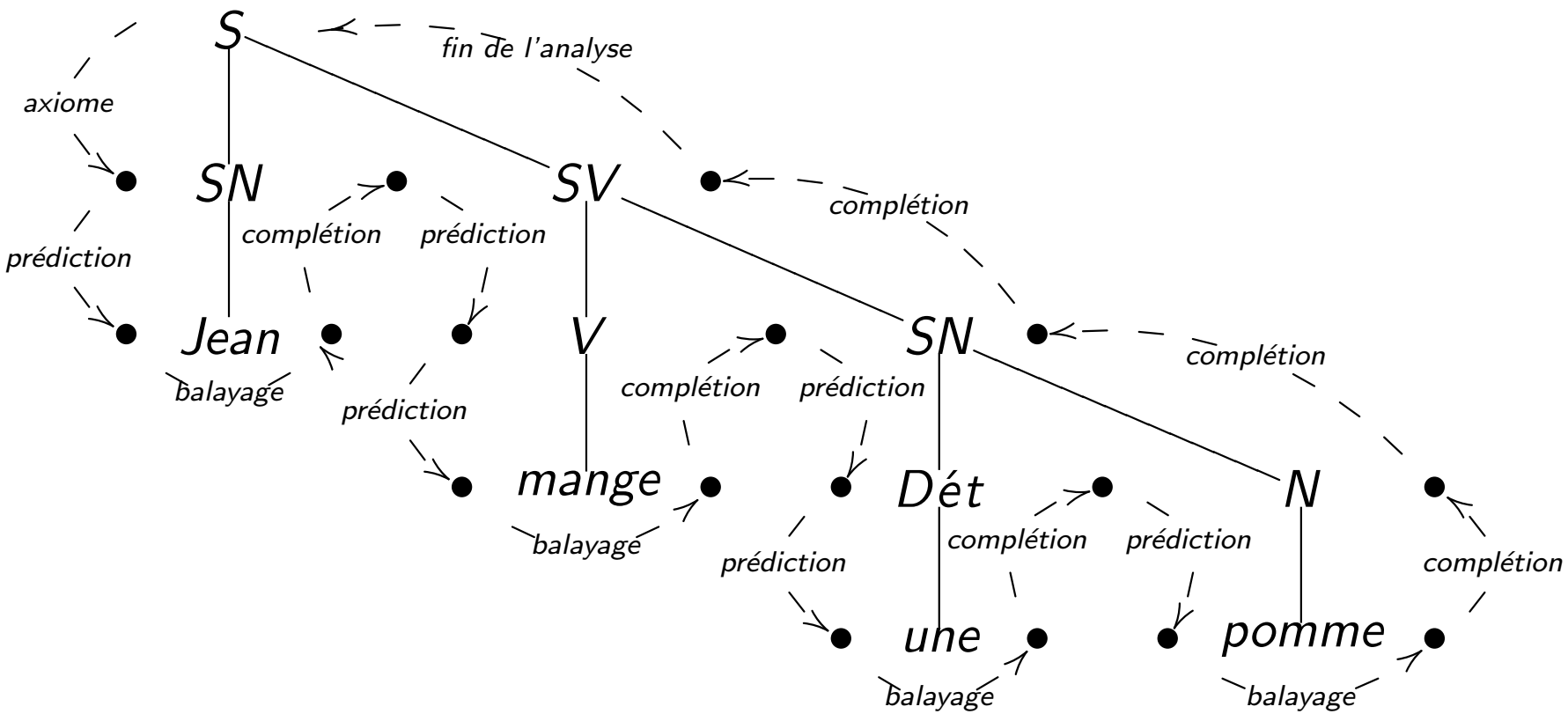
Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Principe



$$
\begin{array}{lcl|lcl}
S & \longrightarrow & SN\ SV & V & \longrightarrow & mange \\
SV & \longrightarrow & V\ SN & Dét & \longrightarrow & une \\
SN & \longrightarrow & Dét\ N \mid Jean & N & \longrightarrow & pomme
\end{array}
$$

FIG.: Arbre d'analyse de Earley pour l'énoncé "Jean mange une pomme"

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Items manipulés par l'algorithme

Les items manipulés par l'algorithme sont de la forme :

$< A \longrightarrow \alpha \bullet \beta, (i, j) >$ représente la situation dans l'analyse où

- $A \longrightarrow \alpha\beta$ est une règle de la grammaire et α est déjà analysé et β est attendu.
- i et j sont les indices représentant la portion de l'énoncé analysé.

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Les règles d'inférence

- Règle axiome

$$\frac{}{< S \longrightarrow \bullet\alpha, (0,0) >}\text{,}$$

pour toute règle $S \longrightarrow \alpha$ de la grammaire avec S l'élément initial.

- Règle de prédiction

$$\frac{< A \longrightarrow \alpha \bullet B\beta, (i,j) >}{< B \longrightarrow \bullet\gamma, (j,j) >}\text{,}$$

pour toute règle $B \longrightarrow \gamma$ de la grammaire.

- Règle de balayage

$$\frac{< A \longrightarrow \alpha \bullet w\beta, (i,j) >}{< A \longrightarrow \alpha w \bullet \beta, (i,j+1) >}\text{,}$$

si w est le mot attendu de l'énoncé.

- Règle de complétion

$$\frac{< A \longrightarrow \alpha \bullet B\beta, (i,j) > \quad < B \longrightarrow \gamma\bullet, (j,k) >}{< A \longrightarrow \alpha B \bullet \beta, (i,k) >}$$

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Principe de tabulation

- Afin que l'analyse soit efficace, il est nécessaire de ne pas analyser plusieurs fois les mêmes sous-arbres.
- Pour cela, les items produits lors de l'analyse sont stockés dans un tableau et les items à traiter sont ordonnancés dans un agenda.

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Introduction

- Les grammaires d'interaction s'appuient sur la notion de *description d'arbres* :
- Une description d'arbres est définie par un ensemble de nœuds et de relations d'ascendance, de parenté et de précédence entre ces nœuds.
- Les nœuds représentent des syntagmes (éventuellement vides) et les relations expriment les dépendances entre ces syntagmes.
- Les propriétés morpho-syntaxiques de ce syntagmes sont décrites par des structures de traits.

Introduction

- L'analyse syntaxique consiste à chercher des modèles de descriptions d'arbres sous forme d'arbres syntaxiques complètement spécifiés.
- Ce processus est hautement indéterministe.
- Limiter cet indéterminisme en contraignant la syntaxe des descriptions et le mécanisme de composition syntaxique.
- Dans la grammaires d'interaction, le mécanisme de composition syntaxique est régi par le principe de *neutralisation de polarités opposées*.
- Certaines ressources munies de polarités négatives sont attendues alors que d'autres, munies de polarités positives, sont disponibles si bien que les premières vont chercher à rencontrer les secondes

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Les descriptions d'arbres polarisées

- Une description d'arbres polarisée est définie par un ensemble de nœuds et par un ensemble de relations entre ces nœuds.
- Chaque nœud est étiqueté d'une structure de traits polarisés. Ainsi, à chaque trait, en plus d'une valeur, est associée une polarité pour indiquer éventuellement si c'est une ressource ou un besoin.
  - $\leftarrow$ indique une polarité positive (ressource)
  - $\rightarrow$ indique une polarité négative (besoin)
  - $=$ indique une polarité neutre
  - $\leftrightarrow$ indique un polarité neutre issue d'une neutralisation
- Les grammaires d'interaction sont lexicalisées. C'est-à-dire que chaque description d'arbres élémentaire est distinguée par son nœud ancre qui exprime la relation entre la description et le lexique.

Les descriptions d'arbres polarisées

- Il existes 4 types de relations :
- Relations de dominance immédiate
  $N_1 > N_2$ signifie que le syntagme N_2 est un constituant immédiat de N_1
- Relations de dominance sous-spécifiée
  $N_1 \overset{*}{>} N_2$ signifie que le syntagme N_2 est inclus dans N_1 à une profondeur indéterminée (éventuellement N_1 s'identifie à N_2)
- Relations de précédence immédiate
  $N_1 \prec N_2$ signifie que le syntagme N_1 précède immédiatement le syntagme N_2 dans l'ordre linéaire des mots de la phrase
- Relations de précédence sous-spécifiée
  $N_1 \overset{*}{\prec} N_2$ signifie que le syntagme N_1 précède le syntagme N_2 dans l'ordre linéaire des mots de la phrase

Les descriptions d'arbres polarisées

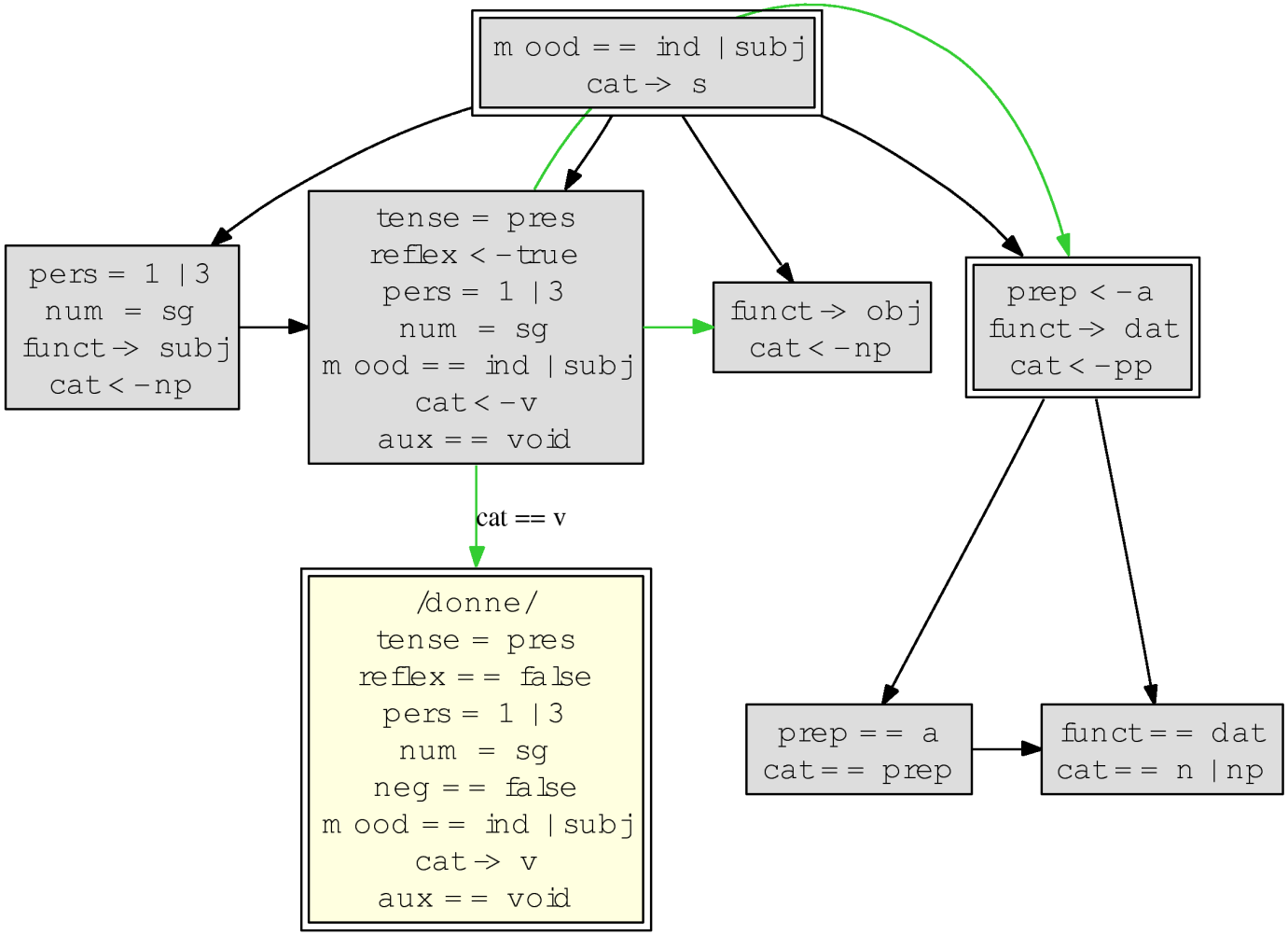


FIG.: Description d'arbres élémentaire du verbe "donne" dans la phrase *qqn donne qqc à qqn*

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Construction de modèles de description d'arbres

- La composition syntaxique de deux descriptions d'arbres est la superposition partielle de ces deux arbres résultant de la fusion de deux nœuds porteurs de traits opposés.
- Analyser une description d'arbres consiste à itérer l'opération de neutralisation des traits opposés pour spécifier progressivement la description initial.
- Cela correspond à la recherche d'un modèle de description d'arbres.
- Un modèle d'une description d'arbres D est une couple formé d'un arbre A et d'une interprétation I :
  - A est un arbre ordonné et ses nœuds sont étiquetés par des structures de traits.
  - I est une fonction d'interprétation de l'ensemble $|D|$ des nœuds de D dans l'ensemble $|A|$ des nœuds de A
- Une analyse réussit si elle s'achève par un arbre complètement spécifié sans relation large où tous les traits ont été neutralisés.

Construction de modèles de description d'arbres

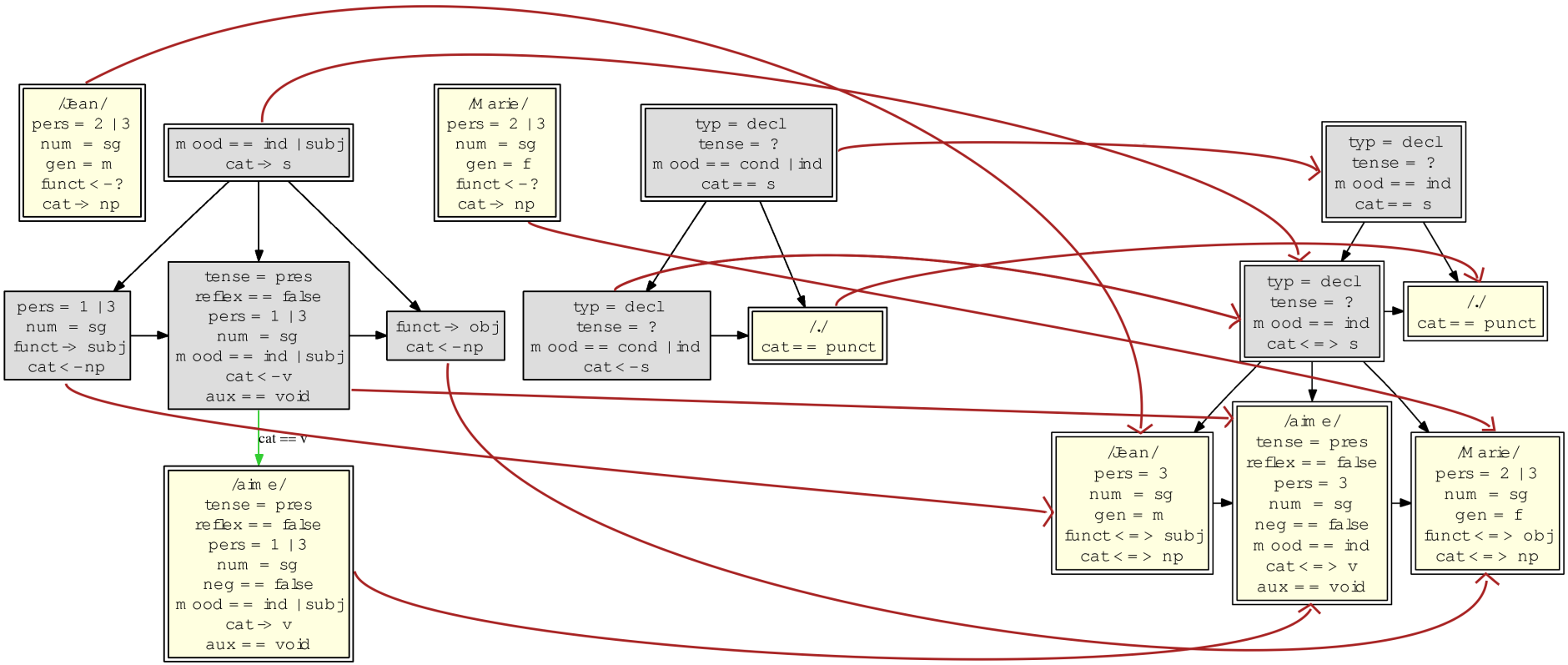


FIG.: Modèle de l'énoncé "Jean aime Marie."

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Intuition

- Cet algorithme reprend en grande partie les travaux de Joseph Le Roux sur un analyseur de Earley pour une version simplifiée des grammaires d'interaction.
- Comme avec l'algorithme de Earley classique, on construit l'arbre d'analyse de haut en bas :
- A partir des racines des descriptions d'arbres élémentaires, l'analyseur spécifie tous les nœuds racines d'un modèle possibles (axiome).
- A chaque étape de la descente, l'analyseur neutralise un nœud du modèle en construction et explore toutes les possibilités de sous-arbres de ce nœud (prédiction).
- A chaque fois que l'analyseur rencontre le mot attendu de l'énoncé, l'analyse avance d'un pas (balayage).
- Quand un sous-arbre est analysé avec succès, l'analyseur remonte dans l'arbre d'analyse et avance d'un pas (complétion).

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Les items manipulés

- Soit $\mathcal{M} : (A, I)$ un modèle de description d'arbres. Un contexte $\{B_1, \ldots, B_n\}$ est l'image inverse d'un nœud B de A par I (ie, $I(B_i) = B$ pour tous les B_i du contexte).
- Les items manipulés lors de l'analyse sont de la forme :

$$< A_{C_A} \longrightarrow \alpha \bullet B_{C_B}\beta, (i, j), (S, U, D) > \text{ où}$$

  - C_A et C_B sont les contextes associés aux nœuds du modèle, ils sont étiquetés par la structure de traits issue de l'unification des structures de traits des nœuds du contexte.
  - $A_{C_A} \longrightarrow \alpha \bullet B_{C_B}\beta$ est une règle pointée, la sémantique de cette règle est la suivante :
    - C_A est complètement spécifié.
    - A est le père de tous les nœuds du corps de la règle pointée et ses fils sont ordonnés de gauche à droite dans le modèle en construction.
  - i et j représentent les indices de la portion de l'énoncé analysé dans la règle pointée.
  - le triplet (S, U, D) représente la situation des ressources de la grammaire à l'étape de l'analyse.

Plan

1. Introduction
2. L'algorithme de Earley
   - Principe
   - Items manipulés par l'algorithme
   - Les règles d'inférence
   - Principe de tabulation
3. Les grammaires d'interaction
   - Introduction
   - Les descriptions d'arbres polarisées
   - Construction de modèles de description d'arbres
4. Un algorithme de Earley pour les grammaires d'interaction
   - Intuition
   - Les items manipulés
   - Les règles d'inférences
5. Conclusions

Les règles d'inférence

- Règle axiome

$$\frac{}{< \top \longrightarrow \bullet S_N, (0,0), (\mathcal{D}, \emptyset, \emptyset) >}$$

- Règle de prédiction

$$\frac{< A \longrightarrow \alpha \bullet B_{\{B_1,\ldots B_n\}}\beta, (i,j), S, U, D >}{< B_{\{B'_1,\ldots B'_{n+m}\}} \longrightarrow \bullet\gamma, (j,j), (S', U', D') >}$$

- Règle de balayage

$$\frac{< A \longrightarrow \alpha \bullet B_{\{B_1,\ldots B_n\}}\beta, (i,j), S, U, D >}{< B_{\{B'_1,\ldots B'_{n+m}\}} \longrightarrow \bullet, (j,j+1), (S', U', D') >}$$

- Règle de complétion

$$\frac{\begin{array}{c}< A \longrightarrow \alpha \bullet B_{\{B_1,\ldots B_n\}}\beta, (i,j), (S, U, D) > \\ < B_{\{B'_1,\ldots B'_{n+m}\}} \longrightarrow \gamma\bullet, (j,k), (S', U', D') >\end{array}}{< A \longrightarrow \alpha B_{\{B_1,\ldots B_{n+m}\}} \bullet \beta, (i,k), (S'', U'', D'') >}$$

Résultats expérimentaux

- Cet algorithme a été implanté dans LEOPAR.
- Cela a permis de comparer cette stratégie aux autres stratégies d'analyses implantées dans LEOPAR et d'obtenir des premiers résultats :
  - La vitesse d'analyse est similaire pour les phrases peu ambiguës mais il y a une baisse de performance quand un grand nombre des racines des descriptions d'arbres peuvent être inclus dans le nœud racine d'un modèle.
  - L'algorithme tabule uniquement sur un étiquetage possible de l'énoncé au lieu de tabuler sur tous les étiquetages. Cela a pour conséquence une chute des performances pour les phrases fortement ambiguës.
  - Un problème lié à une analyse tabulaire est qu'il y a une explosion de l'espace mémoire utilisées par l'analyseur. Cela est dû à la création de trop d'items inutiles lors de la phrase de prédiction.

Pistes d'amélioration

- Introduction d'un symbole initial dans la grammaire pour réduire la combinatoire à la première étape de l'analyse.
- Implantation de la tabulation sur tous les étiquetages d'un énoncé. Cela permettrait d'obtenir des résultats comparables aux autres stratégies d'analyse pour les phrases très ambiguës.
- La grande faiblesse de cet algorithme de Earley est la surgénération d'items lors de la phase de prédiction :
  - Descendre la combinatoire au plus bas dans l'analyse en ne construisant le modèle qu'avec la fusion de nœuds apportant une polarité positive ou négative, et en recollant les les nœuds neutres seulement s'il y en a besoin.
  - Effectuer une analyse de Earley guidée qui permettrait de restreindre le choix des items produits lors de la phase de prédiction.