# Un point sur les outils du LPL pour l'analyse syntaxique du français

Stéphane Rauzy & Philippe Blache

Laboratoire Parole et Langage
CNRS & Université de Provence

stephane.rauzy@lpl-aix.fr
philippe.blache@lpl-aix.fr

# Analyse syntaxique du français

Les modules et ressources développés au LPL pour l'analyse syntaxique du français.

Une approche symbolique ou stochastique, selon les caractéristiques de la tâche à accomplir.

La chaîne de traitement :

- Segmenteur par règles et lexique
- Etiqueteur morphosyntaxique probabiliste
- Deux analyseurs de surface, l'un symbolique et l'autre stochastique
- Un analyseur stochastique profond

# Segmentation et lexique

Un segmenteur par règles :

- Repérage des frontières des tokens
- Identification des entités nécessitant un traitement spécial (nombres, dates, heures, noms propres, sigles, ...)

Un lexique couvrant du français (440 000 formes) régulièrement corrigé et mis à jour. Pour chaque token, la liste des catégories morphosyntaxiques associées et leurs fréquences lexicales.

| graphie | catégorie | fréquence |
|---|---|---|
| ... | ... | ... |
| dans | Nom | 195 |
| dans | Préposition | 1 056 924 |
| ... | ... | ... |
| envers | Nom | 2 123 |
| envers | Préposition | 5 174 |

Information sur les fréquences lexicales : A elle seule, un score de désambiguisation de 0.89 (F-Mesure).

# Etiquetage morphosyntaxique

- Désambiguisation : une catégorie unique pour chaque token
  - Le modèle des patrons (Blache&Rauzy 2006), un HMM plus performant que les N-grammes.
  - Les états de l'automate sont identifiés par des séquences de catégories de longueur variable
  - Apprentissage sur le corpus étiqueté Grace/Multitag (Paroubek et al. 2000), 700 000 mots annotés selon le jeu de traits Multext.
  - Dans l'étude, l'information morphosyntaxique disponible est groupée en 44 catégories distinctes (e.g. les traits d'accords en genre, nombre et personne ne sont pas considérés).
- Pour le jeu de catégories sélectionnées, un score de 0.94 (F-mesure).

# Analyse syntaxique superficielle

L'analyseur superficiel ShP1 :

- Un analyseur symbolique déterministe reposant sur le formalisme des Grammaires de Propriétés avec une stratégie de coin gauche.
- Une grammaire complète pouvant être utilisée indifféremment pour une analyse profonde ou superficielle (Balfourier et al 2005).
- L'analyseur ShP1 s'appuie sur un sous-ensemble de contraintes de la grammaire (en particulier les propriétés de linéarité et de constituance) pour identifier les coins gauches.
- La stratégie consiste à repérer à partir des coins gauches la frontière droite du chunk sur la base des autres propriétés.

# Analyse syntaxique superficielle

L'analyseur stochastique StP1 :

- Modèle des patrons entrainé sur le gold standard Easy (Paroubek et al. 2006, 100 000 mots annotés en constituants). Une phase préalable d'étiquetage du gold standard.
- Environ 1000 patrons de taille variable identifiés par des séquences de catégories terminales (les catégories morphosyntaxiques) ou non-terminales (les groupes Easy).
- Pour chaque énoncé, l'algorithme de Viterbi permet d'insérer les groupes Easy maximisant la probabilité de l'énoncé.

Dans le cadre de la campagne Passage 2007, StP1 et ShP1 ont obtenus des bons scores pour le chunking (resp. 93.03 % et 91.57 %)

# Un analyseur stochastique profond

Extension du modèle des patrons pour le traitement des structures arborescentes.

Les catégories du modèle :

- Les terminales, ex. `Det`, `Verb`, ...
- Les ouvertures, fermetures et feuilles des non-terminales, ex. `<NP>`, `</PP>`, `VP`, ...

Les états (patrons) du système :

- L'identifiant du patron, ex. `NP <VP> Verb`
- Pour chaque catégorie du modèle, ex. `Det` :
  - La probabilité de transition, ex. `p(Det | NP <VP> Verb)`
  - Les catégories insérées, ex. `<NP>`
  - Le patron cible, ex. `<VP> Verb <NP> Det`

# Un analyseur stochastique profond

Hypothèse de réduction : La probabilité de conditionnement par un syntagme ne dépend pas des constituants du syntagme

$\Longrightarrow$ `p(<VP> | <NP> Det Adj Noun </NP>) = p(<VP> | NP)`

C'est une hypothèse forte, il est parfois nécessaire de modifier la définition des syntagmes :

`<SENT><NP>` *Je* `</NP><VP><NP>` *le* `</NP>` *vois* `</VP>` . `</SENT>`

Ajout d'une distinction `NP` - `NPpro` :

`<SENT><NP>` *Je* `</NP><VP><NPpro>` *le* `</NPpro>` *vois* `</VP>` . `</SENT>`

Les patrons du modèle sont obtenus par apprentissage sur un corpus arboré (par exemple le FTB d'Abeillé, le MFT, le LPL Treebank).

# Un analyseur stochastique profond

Calcul de la probabilité d'une séquence arborée :

- C'est le produit des probabilités de transition d'état à état.
- Une procédure de *backtracking* permet d'atteindre le patron cible lorsque une catégorie fermante est insérée (i.e. phase de réduction).

| t | back | patron | insert | cat |
|---|---|---|---|---|
| 0 | - | <SENT> | <NP> | Det |
| 1 | - | <SENT><NP>Det | - | Noun |
| 2 | - | <NP>Det Noun | </NP> | - |
| 3 | 0 | <SENT>NP | <VP> | Verb |
| 4 | - | NP<VP>Verb | <NP> | Det |
| 5 | - | Verb<NP>Det | - | Noun |
| 6 | - | <NP>Det Noun | </NP> | - |
| 7 | 4 | NP<VP>Verb NP | </VP> | - |
| 8 | 3 | <SENT>NP VP | - | Pct |
| 9 | - | <SENT>NP VP Pct | </SENT> | - |

# Un analyseur stochastique profond

Le parsing (calcul de la ou des meilleures solutions) est effectué par un algorithme de type *beam search*.

La grammaire probabiliste générée fait partie de la classe des grammaires sensibles au contexte (*Context Sensitive Grammar*).

Application : Apprentissage sur *Modified French Treebank*.

- Des résultats préliminaires très encourageants.

Le traitement des relations :

- Pour le MFT, les relations sont annotées en précisant la fonction du syntagme (ex. `SUBJ`, `OBJ`, `MOD`, ...).
- Cette information peut être directement incorporée dans l'analyseur, en sous-catégorisant les syntagmes existants (e.g. on distinguera les catégories `NP:SUBJ`, `NP:OBJ`, `NP:MOD`, ...).
- Quel est le gain (résolution des attachements, ...) ?

# Conclusion et perspectives

Une chaîne de traitement complète pour l'analyse syntaxique du français.

- Des analyseurs de surface (stochastique et symbolique).
- Un analyseur stochastique profond.

Travail en cours :

- Evaluation des résultats.
- Spécification et enrichissement du corpus d'apprentissage (catégorisation, phénomènes traités, taille).
- Amélioration du modèle (traitement de l'accord, ...).