Analyse linguistique automatique

Mathieu Constant
Université de Lorraine, CNRS, ATILF

Journée IA, Langage et Citoyens
15 mars 2019

Aujourd'hui

- À travers un retour historique, réfléchir sur l'évolution de l'analyse linguistique automatique
- Interprétabilité linguistique des résultats ?
- Contrôle des outils ?
- Compétences ?

Projet IMPACT OLKi

Groupes de travail

- Plateforme fédérée
- Connaissance et ingénierie
- Langage
- Défis sociétaux

Axes transversaux

- Aspects éthiques
- Apprentissage automatique

Projet IMPACT OLKi

Groupes de travail

- Plateforme fédérée
- Connaissance et ingénierie
- Langage
- Défis sociétaux

Axes transversaux

- Aspects éthiques
- Apprentissage automatique

Projet IMPACT OLKi II

Groupe de travail Langage

- Ressources linguistiques
- Traitement automatique des langues
- Apprentissage (humain) des langues

Projet IMPACT OLKi II

Groupe de travail Langage

- Ressources linguistiques
- Traitement automatique des langues
- Apprentissage (humain) des langues

Traitement automatique des langues

Applications

- traduction automatique
- résumé automatique
- systèmes questions-réponses
- extraction d'information, ...

Une discipline historiquement liée à la linguistique

- **Linguistique**: décortiquer et comprendre les mécanismes de la langue, notamment la construction du sens
- **Différents niveaux d'analyse**: phonologie, morphologie, syntaxe, sémantique, pragmatique, ...

Traitement automatique des langues

Applications

- traduction automatique
- résumé automatique
- systèmes questions-réponses
- extraction d'information, ...

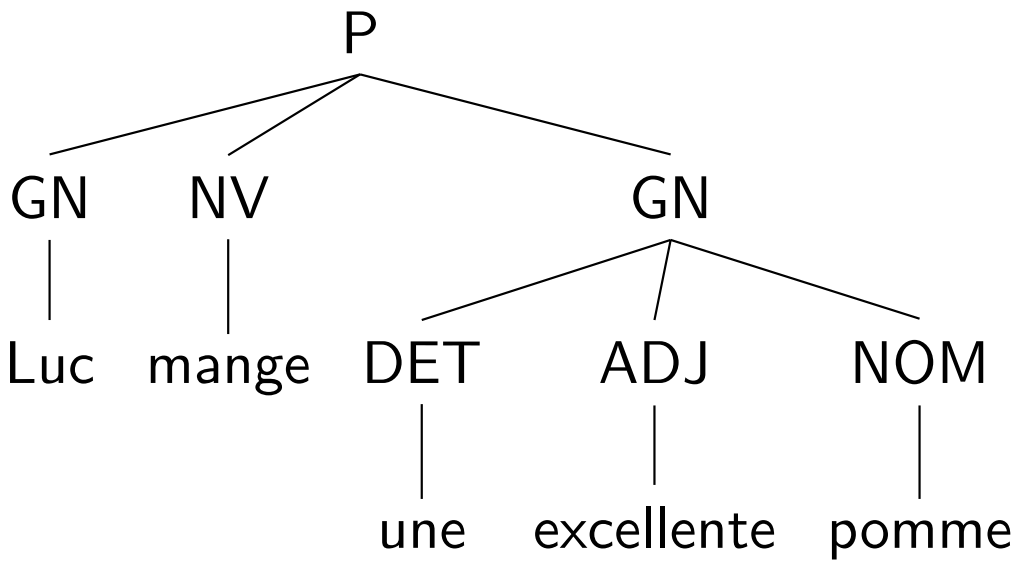
Une discipline historiquement liée à la linguistique

- **Linguistique**: décortiquer et comprendre les mécanismes de la langue, notamment la construction du sens
- **Différents niveaux d'analyse**: phonologie, morphologie, syntaxe, sémantique, pragmatique, ...

Analyse linguistique automatique

- ex1. Etiquetage grammatical des mots
- ex2. Prédire la structure syntaxique d'une phrase,
- ex3. Construire la représentation du sens des mots et d'une phrase
- **Défis**: ambiguïté naturelle de la langue, modélisation du sens et des mécanismes de composition, ...

Exemple : analyse syntaxique en constituants



Évolution des approches d'analyse syntaxique automatique I

- Outils de plus en plus performants et robustes
- En contrepartie, une perte de "contrôle" et d'interprétabilité progressive sur les outils d'analyse
- En termes de compétences, on avait des linguistes informaticiens, maintenant on a des ingénieurs informaticiens

Approches fondées sur des formalismes grammaticaux I

À partir des années 60 - 70

Approche

- Développement de formalismes linguistiques
- Développement de grammaires et d'outils s'appuyant sur ces formalismes
- Extensions vers analyse sémantique et discursive
- En parallèle, développement de ressources lexicales

Approches fondées sur des formalismes grammaticaux II

À partir des années 60 - 70

Exemple : Context-free grammars

$P \rightarrow GN\ NV\ GN$

$GN \rightarrow NOM$

$GN \rightarrow DET\ NOM$

$GN \rightarrow DET\ ADJ\ NOM$

$NV \rightarrow V$

...

Exemple : tree adjoining grammar, extrait de (Joshi et Schabes 1997)

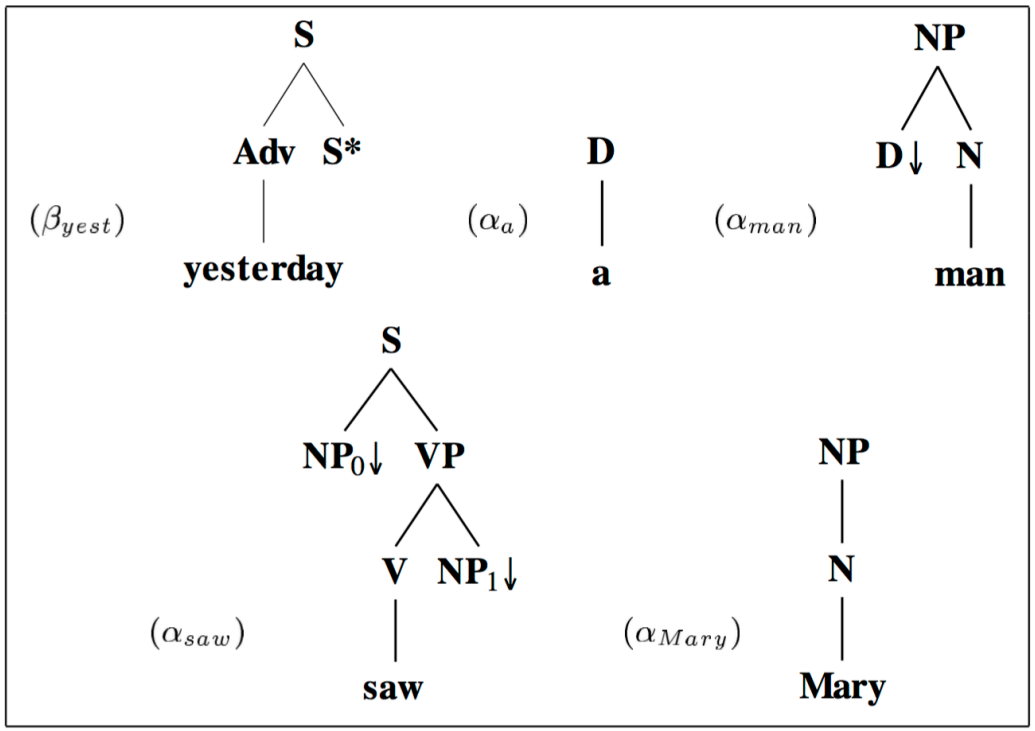


Fig. 2.4. Some elementary trees.

Approches fondées sur des formalismes grammaticaux IV

À partir des années 60 - 70

Quelques conclusions

- Interprétabilité linguistique des résultats
- Développement : besoin d'une expertise linguistique et formalismes, algorithmes d'analyse
- Développement monolingue

Intégration de modèles statistiques

À partir des années 90

- Problème de l'ambiguïté : les grammaires sont accompagnées de modèles statistiques
- Ex. assigner des probabilités aux règles du formalisme grammatical sous-jacent
- Développement de corpus annotés : Penn Treebank (Markus et al. 1993) $\rightarrow$ extraction de grammaires probabilistes

Analyseurs sans formalismes grammaticaux I

A partir des années 2000

Exemple : construction incrémentale de la structure syntaxique

- prédiction d'une suite d'opérations élémentaires sur base de décisions locales (ex. Nivre 2003)
- décisions locales prises par un classifieur appris sur corpus annoté

Feature engineering

- Les décisions locales sont basées sur des traits linguistiques
- C'est le développeur du système qui décide de l'ensemble des traits utilisés
- L'optimisation du système passe par une longue phase d'ajustement des traits utilisés

Exemple de traits (Zhang et Nivre 2011)

from single words
S_0wp; S_0w; S_0p; N_0wp; N_0w; N_0p; N_1wp; N_1w; N_1p; N_2wp; N_2w; N_2p;
from word pairs
S_0wpN_0wp; S_0wpN_0w; S_0wN_0wp; S_0wpN_0p; S_0pN_0wp; S_0wN_0w; S_0pN_0p N_0pN_1p
from three words
$N_0pN_1pN_2p$; $S_0pN_0pN_1p$; $S_{0h}pS_0pN_0p$; $S_0pS_{0l}pN_0p$; $S_0pS_{0r}pN_0p$; $S_0pN_0pN_{0l}p$

Table 1: Baseline feature templates.
w – word; p – POS-tag.

Analyseurs sans formalismes grammaticaux III

A partir des années 2000

Conclusions

- Interprétabilité linguistique des résultats : correlation entre poids des traits et résultats obtenus
- Développement : algorithme d'analyse, modèle statistique, définitions des traits (sur intuitions linguistiques et retours d'expériences)
- Outil générique, mais les traits sont parfois différents selon les langues

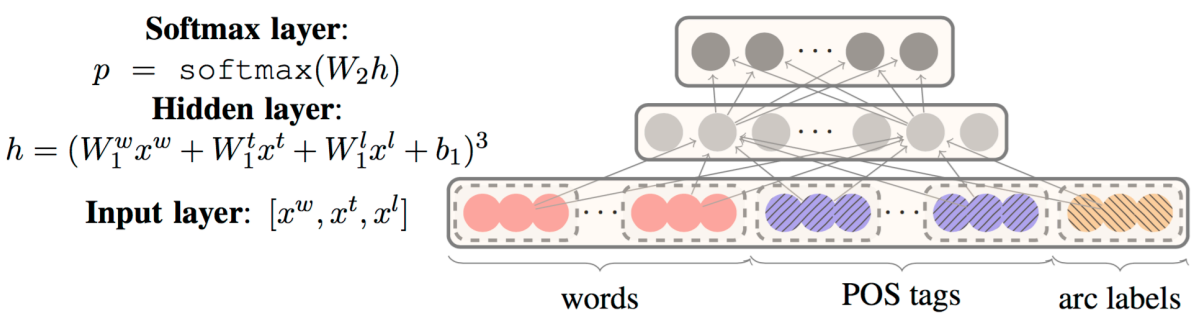
Passage à l'apprentissage profond I

Dans les années 2010

Deux avancées significatives

- plongements de mots (word embedding) : représentation (sémantique) des mots dans un espace continu de dimension réduite (Mikolov et al. 2013)
- le renouveau des réseaux de neurones : Collobert et al. (2011), Chen et Manning (2014) pour l'analyse automatique

Exemple pour l'analyse syntaxique (Chen and Manning 2014)



Passage à l'apprentissage profond II

Dans les années 2010

Quelques caractéristiques

- Les entrées du réseau : des vecteurs représentant des éléments linguistiques (caractères, mots, catégories grammaticales, ...)
- Plus besoin de feature engineering !
- Par contre, besoin d'ajuster de nombreux d'hyperparamètres à ajuster (dimensions des vecteurs, nombres de couches, learning rate, ...). Cauchemar pour certaines tâches !

Passage à l'apprentissage profond III

Dans les années 2010

Quelques conclusions

- Interprétation linguistique des résultats très difficile
- Approche générique et multilingue
- Compétences : apprentissage automatique, quelques notions linguistiques pour donner les bons éléments en entrée

La linguistique pour le TAL n'est pas morte ... I

Production de données annotées

- Pour réaliser des tâches d'analyse linguistique, besoin de beaucoup de données annotées linguistiquement
- Des initiatives internationales collectives pour constituer des **corpus annotés libres**
  - analyse syntaxique : Universal Dependencies, plus de 100 corpus pour 70 langues (Nivre et al. 2014)
  - reconnaissance d'expressions verbales : 18 langues (Savary et al. 2017)
  - ...
- Par contre, ces initiatives sont guidées par les besoins du TAL : normalisation, appauvrissement de la description linguistique

La linguistique pour le TAL n'est pas morte ... II

Utilité des ressources linguistiques

- Les ressources lexicales ont leur rôle à jouer dans l'interprétabilité des résultats
- Il existe des plateformes de dépôts de ressources linguistiques extrêmement riches (ORTOLANG , CLARIN)
- La quantité, la finesse et l'hétérogénéité des ressources linguistiques sont une richesse. C'est à l'intelligence artificielle de s'adapter et pas uniquement l'inverse.

Conclusion

- Apprentissage profond : accroissement des performances, démocratisation des outils (approches génériques), multilinguisme, moins d'interprétabilité, moins de contrôle
- La part de la linguistique est maintenant limitée à la production de données, de manière encadrée
- De nombreuses ressources linguistiques très riches restent inexploitées. L'intelligence artificielle y a beaucoup à gagner.