Des bigrammes au secours de la grammaire

Eric Wehrli, Luka Nerima et Meghdad Farahmand

LATL-CUI Université de Genève
http://latl.unige.ch

Présenté par Yves Scherrer – ALPAGE

Le problème

- Fips, comme tous les analyseurs syntaxiques basés sur des règles de grammaire, surgénère.
- La surgénération provient de la combinatoire qui s'exerce du fait de l'ambiguïté lexicale et de la pluralité des règles d'attachement.
- Des heuristiques sont donc nécessaires pour limiter le nombre d'alternatives et rendre le système utilisable dans des applications pratiques.
- L'élagage de l'ensemble d'alternatives se fait après l'attachement d'un nouveau mot sur la base des scores attribués à chaque analyse et en fonction du spectre choisi (nombre d'analyse maximal, habituellement entre 40 et 80).

Analyse Fips

- Résultats obtenus en fonction du nombre d'alternatives
  (corpus de l'hebdomadaire The Economist 2012, env. 75'000 mots)

max. alternatives	temps	analyses complètes	mots/secondes
120	21 :06	77.98%	59
80	13 :28	75.45%	93
40	6 :26	71.80%	198
10	2 :17	59.40%	579
5	1 :39	50.94%	793

- Ces données montrent que l'élagage permet effectivement d'améliorer considérablement l'efficacité de l'analyseur (quasi-linéaire), mais au détriment du niveau de qualité, estimé sur la base du nombre d'analyses complètes obtenues.

Analyse complète

- Par analyse complète on entend une structure syntaxique bien formée couvrant l'intégralité de la phrase.
- L'analyse complète est un critère de qualité approximatif (complet ne signifie pas nécessairement correct !).
- Il existe pourtant une forte corrélation entre analyse complète et niveau de qualité (découpage lexical et structure syntaxique)

Objectifs

- Améliorer l'élagage de manière à réduire (idéalement éliminer) la perte qualitative actuelle.
- Remplacer les scores «manuels» actuels par des probabilités statistiques calculées sur de gros corpus de textes.
- Questions :
  - Quels types de données statistiques
    - bigrammes de mots
    - bigrammes de lexèmes
    - bigrammes de **mots+catégorie**
  - À quel moment de l'analyse convient-il de les utiliser
    - avant l'analyse syntaxique, lors d'un prétraitement lexical?
    - **pendant l'analyse syntaxique**

Première expérience

- Utiliser les bigrammes pour favoriser les lectures les plus probables de mots homographes.

  - Exemples d'homographes

    [nom – verbe conjugué] démarche, options, notions, menace, affluent, souris, président, …

    [nom – participe] correspondant, étudiant, représentant, élu, initié

    [conjonction – préposition] à, pour, au lieu de, …

    [nom – adjectif] passager, français, patient, …

    …

  - Quelques exemples anglais:

    [nom – verbe] record, cost, make, mix, …

Première expérience (suite)

- Au moment où l'analyseur cherche à combiner un terme homographe, pour chacune des possibilités de combinaison, on consulte les bigrammes pour les 2 termes afin d'attribuer un score à la structure résultant de cet attachement.
- A l'issue du traitement pour un mot donné dans la phrase, on filtre l'ensemble des structures sur la base de leur score.

Acquisition des bigrammes

- Extraction des bigrammes
  - Bigrammes de mots avec catégorie
    [marketing N, costs N], [marketing N, costs V]
  - On s'intéresse uniquement aux bigrammes t.q. le deuxième mot est un homographe (|catégorie lexicale| > 1)
  - La liste des homographes est obtenue par interrogation de notre base de données lexicale des formes fléchies :

    7'022 en français (sur approx. 250'000 formes)
    9,675 en anglais (sur approx. 100'000 formes)

Corpus

- Corpus
  - Extrait de Wikipedia anglais, ~ 140 Mio de mots
- Quelques chiffres
  - Taille du corpus anglais : 139’690’263 mots
    nombre d’occurrences d’homographes 37’341’534 (soit 26 %)
  - Taille du corpus français : 82’896’739 mots
    nombre d’occurrences d’homographes 28’969’353 (soit 34 %)
- Prétraitement
  - Étiquetage POS avec le Stanford Parser (Klein et Manning 2003)

Calcul de la probabilité

- Bigramme: (préfixe, homographe)
- Notation: (*Pref*, *H*)

C : Count

POS_i^H : the i^th POS tag of Homograph H

$Pref$: Homograph's prefix

POS^{PREF} : POS tag of the prefix

S : set of i^{th} homograph POS tags (for a given $Pref$ and POS^{PREF})

$$P(POS_i^H | Pref, POS^{PREF}) = \frac{C(POS_i^H \text{ follows } Pref \text{ and } POS^{PREF})}{\sum_{j \in S} C(POS_j^H \text{ that follow } Pref \text{ and } POS^{PREF})}$$

- Exemple (Wikipedia)

$P(\text{N}^{\text{costs}} \mid \text{N}^{\text{marketing}}) = 0.92$

$P(\text{V}^{\text{costs}} \mid \text{N}^{\text{marketing}}) = 0.08$

Etiquettes Stanford → étiquette Fips

- Conversion
  - Stanford: 36 étiquettes POS pour l'anglais (Penn Treebank)
  - Fips: 9 étiquettes
  - Par exemple

    CC Coordinating conjunction → Conj

    CD Cardinal number → Det

    DT Determiner → Det

    etc
- Problème
  - Pas de correspondance directe entre les jeux d'étiquettes
  - Par exemple:

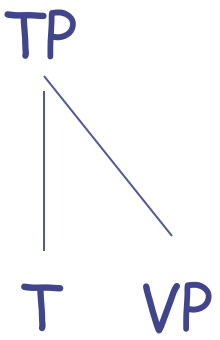
    l'étiquette IN "Preposition or subordinating conjunction" → Prep ou Conj ?

Fonction de probabilité pour Fips

- Fonction
  - ((Préfixe,catégorie), (Homographe, catégorie)) → probabilité
- Optimisation
  - Utilisation des index de formes fléchies de la base lexicale de Fips:
    (Préfixe,catégorie) → index de mot du préfixe
    (Homographe, catégorie) → index de mot de l'homographe
- Implémentation
  - Concaténation des deux index, hachage
  - Tableau à une entrée → accès *O*(1)

Rappel: Les structures syntaxiques de Fips

- Dans Fips, les structures syntaxiques sont créées par projection
  - soit à partir d'un item lexical:
    X --> XP X ∈ {N, V, A, D, P, Adv, Conj}
  - projection étendue (métaprojection):
    verbe conjugué -->



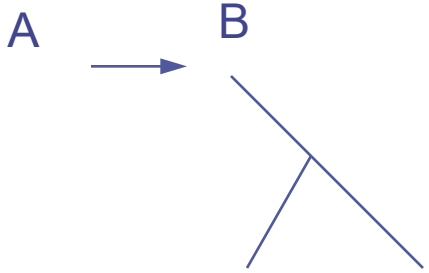
pronoms et noms propres --> DP

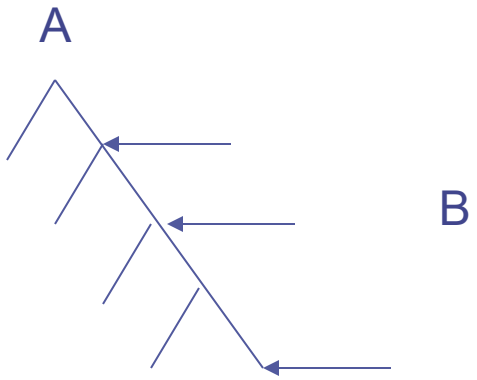
Utilisation de la fonction

- La fonction est invoquée lors de l'attachement d'un terme homographe
- Fips considère des attachements à gauche (attachement du sujet, d'un adjectif prénominal, etc.) et à droite (tous les compléments, etc.)
- Les attachements à droite peuvent se faire à plusieurs niveau de la structure (transparent suivant)
  → en cas d'attachement à un nœud supérieur au coin droit, le préfixe et l'homographe ne sont pas adjacents
  - Exemple: … ne **porte** pas toujours **atteinte** …

Attachement de constituants

- L'algorithme d'analyse combine deux projections adjacentes A et B
  - attacher A comme sous-constituant gauche de B (=attachement à gauche) ;
  - attacher B comme sous-constituant d'un nœud actif sur l'arête droite de A (attachement à droite) :





Exemple

(1) Unit labour **costs** measure the average cost of labour per unit of output.
les coûts unitaires de la main-d'oeuvre mesurent le coût moyen de la main-d'oeuvre par unité produite.

- Dans l'exemple ci-dessus, lorsque l'analyseur lit le mot homographe *costs* (substantif ou verbe), il consulte la table des bigrammes pour chacune des deux catégories (N et V), étant donné le verbe *labour*, dans une première analyse, le substantif *labour* dans une deuxième analyse.
- La probabilité associée au bigramme [labour-N, costs-N] (*coût de la main-d'oeuvre*) est largement supérieure à celle des autres bigrammes. Cette analyse sera donc privilégiée.

Questions ouvertes

- Granularité du préfixe
  - Mot (forme fléchie) ?
  - Lexème ?
  - Catégorie lexicale (POS) ?
- Pertinence de ne s'intéresser qu'aux homographes ?
- Calculer des bigrammes plus « sophistiqués », de têtes sémantiques de syntagmes ?

  [Le gros **chat** noir] [a **mangé** [une petite **huître**]]

- Remarque

  Notre approche se distingue d'un étiquetage POS en prétraitement, par le fait que les lectures non privilégiées sont conservées jusqu'à un élagage éventuel de l'analyse.