Approches catégoriques et non catégoriques en linguistique des corpus spécialisés

Application à un système de filtrage d'information

Antonio BALVET, UMR MoDyCO
Paris X

Plan

1. Pour une linguistique des corpus
2. Recherche d'information et linguistique appliquée
3. Filtrage d'information
4. Le système CORAIL
5. Synthèse et perspectives

1. Pour une linguistique des corpus

- Linguistique appliquée:
  - Ingénierie linguistique, enseignement des langues
  - Théorie linguistique?

- Influence du modèle classique des catégories
  - Modèles monocatégoriels, catégories étanches

- Le statut des observations empiriques
  - Chomsky: empirique = non scientifique
  - Herdan, Manning, Pereira, Abney: empirique = statistique = scientifique (Théorie de l'Optimalité)

$\Rightarrow$ Pour une linguistique de corpus non catégorique

2. Recherche d'information et linguistique de corpus

- La langue naturelle effective comme objet d'étude
  - Succès des analyses syntaxiques locales, de surface
  - Échec des approches linguistiques exclusivement catégoriques

- Deux dimensions
  - Dimension subjective: fonction informative (stratégie individuelle)
  - Dimension objective: visée normative, à usage collectif (indexation)

- Les signatures thématiques (Riloff, 1993):
  - Unités lexicales complexes associées à un thème, composées d'éléments lexicaux fortement cohésifs
  - Détection par analyse distributionnelle classique et par approches statistiques (collocations)

3. Filtrage d'information

- Sous-domaine de la recherche d'information
  - Spécifications TREC (Text REtrieval Conference)
    - Décision de sélection binaire, besoin en information stable, fréquence élevée de la mise à jour du fonds documentaire
  - Bilan critique de TREC
    - Les données de référence ne sont pas issues d'une pratique effective
    - Les corpus ne sont pas à taille humaine (10 Go)
    - Métriques inadaptées: utiliser précision, rappel, corrélation réponses/cible

4. CORAIL

- Un système de filtrage d'information
  - Sélection de documents par reconnaissance de signatures thématiques (INTEX)
    - Mise en œuvre de connaissances linguistiques
    - Familles de grammaires locales
  - Utilisabilité évaluée sur des utilisateurs non informaticiens et non linguistes
    - Les grammaires locales comme filtres
    - Développer des assistants linguistiques (utilisateurs, concepteurs de ressources lexicales): LIZARD
  - Efficacité des analyses locales
    - Bonne corrélation signatures thématique/référence (finance)

4.1 Un exemple de signatures thématiques (finance)

Dassault Systèmes acquiert SRAC.
Le groupe met la main sur une société américaine de CAE.
Dassault Systèmes vient de réaliser une nouvelle acquisition : l'un des fleurons de la famille Dassault va mettre la main sur la société américaine SRAC dans une transaction de 22 millions de dollars en actions.

Famille:

- **N0 acheter N1**, actif (*Dassault acquiert SRAC*), nominalisation avec verbe support (*DS vient de réaliser une acquisition)*

4.1 Un exemple de signatures thématiques (finance)

Dassault Systèmes acquiert SRAC.
Le groupe met la main sur une société américaine de CAE.
Dassault Systèmes vient de réaliser une nouvelle acquisition : l'un des fleurons de la famille Dassault va mettre la main sur la société américaine SRAC dans une transaction de 22 millions de dollars en actions.

Famille:

- **N0 Const N1** (*Le groupe met la main sur une société … de CAE, Dassault va mettre la main sur … SRAC*)

4.2 LIZARD

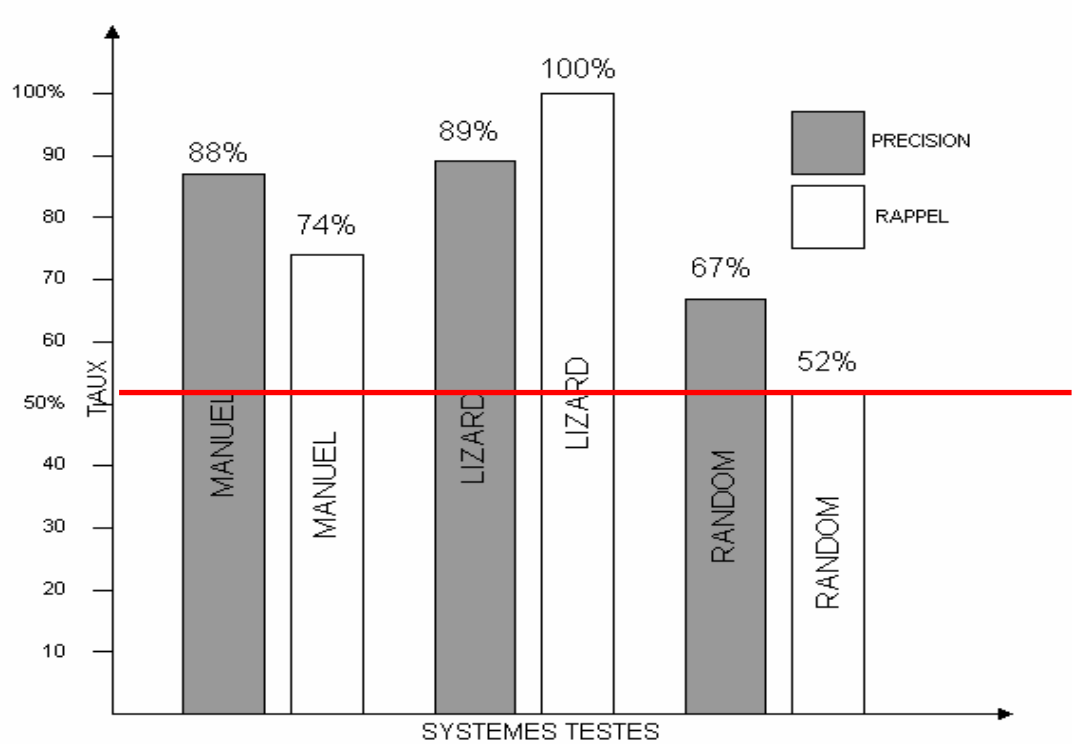
- Linguistic wizard
  - Assistant linguistique pour l'élaboration de ressources lexicales
    - Analyse distributionnelle classique
    - Recyclage d'étiquettes morphosyntaxiques
    - Procédures d'approximation (Harris, 1951)
      - Généralisation:

        acheter la société => acheter DET société => V DET N

        (N0 V N1) ⇔ acquérir DET N
    - Consultation de ressources externes (Memodata)

⇒ Constituer des bases de signatures thématiques

4.3 Évaluation quantitative



Rappel: Bonnes réponses / Réponses fournies
Précision: Bonnes réponses / Réponses attendues
Scores de Khi2 pour Manuel et Lizard: bonne corrélation réponses données / référence

5. Synthèse et perspectives (1/2)

- Linguistique de corpus et recherche d'information
  - L'apport des analyses locales, peu profondes
    - (Riloff, 1993): extraction d'unités lexicales complexes, mi-chemin entre analyse profonde et approches à base de mots-clés
    - Bonne corrélation signatures / thèmes
    - Tester sur d'autres corpus, d'autres moteurs d'analyse (FSM, Unitex…)
  - Perspectives pour les approches non catégoriques
    - Apprentissage d'une grammaire locale d'un domaine de spécialité (Charniak, 1993)
    - Rendre compte de la variation dans la décision de sélection humaine

5. Synthèse et perspectives (2/2)

- Un cadre pour une linguistique de corpus
  - Théorie de l'Optimalité (OT), (Prince & Smolensy, 1993)
    - Satisfaction de contraintes multiples et contradictoires
    - OT statistique: gradient de grammaticalité (Boersma & Hayes, 2001)
    - Applications en syntaxe: grammaire des usages effectifs (Manning, 2002)
    - Saussure/Herdan: la Langue comme tendance, comportement moyen
  - Point de vue objectif/subjectif
    - Décrire la Parole (subjectif) en visant la Langue (objectif) dans un cadre formel
    - Point de vue scientifique non catégorique et non nécessairement logique