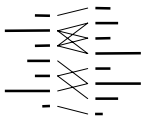
Xerox Research Centre Europe
xrce
A Xerox Research & Technology Organization

Applications multilingues à XRCE

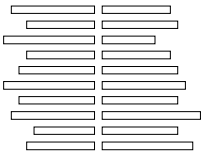
Cyril Goutte

Cyril.Goutte@xrce.xerox.com

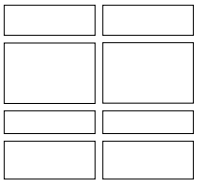
From aligned to parallel to comparable corpora



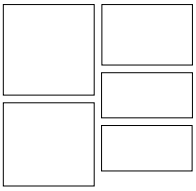
Word-aligned



Aligned



Parallel



Comparable

Corpus aligné

Alignements de mots:

- Projection d'annotation d'une langue à l'autre
- Traduction d'unités sous-phrastiques

Alignements de phrases:

- Modèles statistiques de traduction (IBM1-5)
- Concordances bilingues

Extraction et enrichissement de lexique

À partir de corpus parallèle:

- techniques bien connues et assez efficaces

À partir de corpus comparables:

- enrichissement d'un lexique existant (ex. langue courante)
- co-occurrences avec les entrées du lexique
- méthodes standard: projections et produits scalaires

Analyse: du monolingue au bilingue

Corpus monolingue:

- LSA / Factor Analysis: modéliser la variabilité des données (eg fréquences de mots)

Corpus bilingue:

- KCCA: modéliser les corrélations entre données bilingues
- Corrélations entre co-occurrences avec entrées du lexique

En coulisses…

LSA: décomposition en valeurs propres

KCCA: problème aux valeurs propres généralisées

$$\mathbf{B}\,x = \lambda\,\mathbf{D}\,x$$

PLSA: décomposition en facteurs “non-négatifs”