Formation générale et formation professionnelle en TAL

Benoît Habert
LIR – LIMSI – CNRS & université Paris X – Nanterre
habert@limsi.fr
http ://www.limsi.fr/Individu/habert/

Tant de TAL !

R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2003

- 680 p. ;
- 38 chapitres ($\approx$ 20 p.) avec références spécifiques ;
- 50 auteur(e)s (30% industrie, 60% anglais et américains) ;
- Glossaire de 600 entrées ;
- 3 volets :
  - niveaux de l'analyse ;
  - étapes de traitement ;
  - domaines d'application : traduction automatique, recherche d'information, extraction d'information, fouille de données textuelles. . .

Faire des liens

analyse linguistique	réduction de termes complexes
méthodes informatiques	transducteurs à états finis
applications	extraction d'information

Quand on parle de *thermiques*$_1$, les yeux de tout parapentiste normalement constitué se mettent à briller [...] L'*ascendance thermique*$_1$ est en effet un des principaux vecteurs d'altitude pour le parapente [...] *Un « thermique*$_1$ *», c'est* avant tout une zone où l'air est plus chaud que l'ensemble de la masse d'air environnante, ce qui lui permet par différence de densité de s'élever au sein de cette masse d'air.

Cité dans [Jacques 03]

Dépasser les variations de surface

- « Grammaires locales » : hyperonymie [Morin 99], définition [Rebeyrolles 00], antonomase [Leroy 00], réduction de termes [Jacques 00]…
- Expressions régulières : outils Unix (`grep`, `sed`) ; langages de script (Perl, Python, PHP) ; langages de programmation (Java 1.4)…
- Extraction d'information : réponse à des questions factuelles ; anonymisation, repérage des personnes dans une page Web (noms propres)…

Composer (avec) les applications

• L'âge des produits-étagère : étiqueteurs, moteurs de recherche, boîtes à outils statistiques…

• Réaliser $\equiv$ composer

• Exemple : réponse à des questions factuelles $\longrightarrow$

– classer la question selon le type de réponse attendue (*Quel est le premier président de la 5ème république ?*) ;
– transformer la question en requête ;
– utiliser un moteur de recherche pour récupérer les documents pertinents ;
– extraire la bribe-réponse.

Apprendre à voir de très près / très loin

• Mettre en place des « truchements » : adapter le format d'entrée d'un module au format d'entrée du suivant

– non structuré (texte de pages Web)
– peu/semi-structuré (Cordial)
– structuré (sorties XML de l'analyseur XIP de Xerox)

• Comprendre ce qui sous-tend un module (forces/faiblesses/adéquation à la tâche)

– Etiquetage Cordial | TreeTagger | WinBrill
– Lemmatisation Cordial | racinisation Snowball [Porter] // pages Web

Apprendre à remodeler/articuler/dominer

- Ancrage dans une linguistique descriptive « outillée »
- Pratique continue de techniques de remodelage :
  - développer un « oeil »
  - choisir les bons outils : transducteurs réguliers, feuilles de style XSLT et transformation d'une page Web
- Connaître les méthologies et savoirs sous-jacents : moteur de recherche / classification / apprentissage artificiel

Aujourd'hui déjà hier

- Rapidité d'évolution du TAL : ATN en 80, grammaires d'unification en 90, approches statistiques en 00
- Coller au présent : le meilleur moyen de vieillir mal
- Se former à un domaine et pouvoir suivre son évolution (/ se former à un métier, à une profession)

La meilleure formation professionnelle : une formation générale ?