Traitement Automatique des Langues et Apprentissage des Langues Assisté par Ordinateur: bilan, résultats et perspectives

Sébastien L'HAIRE

Soutenance de thèse

15 juin 2011

Université de Genève

Plan

- Apprentissage des langues
- Traitement Automatique des Langues
- Réalisations présentées dans la thèse
- Correction orthographique
- FipsOrtho
- Conclusions

Apprentissage des langues

- Apprendre une langue: besoin ou plaisir
- Années 1970-80: béhaviorisme,
  répéter des schémas
- Langue → communiquer, accent sur oral
- Tendance: la forme est moins importante (sauf écrit)
- Apprenant (rôle actif) plutôt qu'élève / étudiant

Apprentissage des Langues Assisté par Ordinateur

- Ordinateur comme outil principal ou complément
  - Classe, centre documentation, domicile, déplacement
  - En groupe ou individuel
- Pour les enseignants, analyse détaillée des résultats / parcours
- Pour les apprenants
  - Accessibilité des outils
  - Moyens variés de présentation
  - Rythme individuel, moins d'inhibition
  - Fiabilité pour certaines corrections
  - Rétroaction immédiate
- La langue doit être manipulée
    - QCM et textes à trous insuffisants pour l'évaluation → besoin d'écrire des phrases complètes

Traitement Automatique des Langues

- Branche de l'Intelligence Artificielle
- Traiter l'oral commme l'écrit, production et compréhension
- Aides intelligentes à l'apprentissage
  - Aide prononciation (reconnaissance vocale)
  - Synthèse vocale: prononcer des textes et mots
  - Outils morphologiques: conjugueurs, déclineurs, analyseurs morphologiques
  - Outils de recherche de corpus
  - Correction orthographique et grammaticale

Réalisations de la thèse

- Recherches autour du projet européen FreeText (2000-2003)
- Interfaces d'aide à l'apprentissage
  - Grammaire en couleur
  - Diagnostic d'erreurs
  - ...
- Correction “sémantique”
- Correction orthographique

Correction orthographique

- Correction orthographique: 1ère étape évaluation écrit
- Reconnaître les mots inconnus et proposer des corrections, les meilleures en premier
- Adaptation aux apprenants de langues étrangères (L2)
- FipsOrtho → extension du prototype FipsCorr (Ndiaye & Vandeventer Faltin 2003, 2004)
  - Correcteur sur le web
  - Soumissions récoltées dans un corpus d'erreurs et évaluées par un expert

Les travails* sont difficiles

- Analyse syntaxique
- Système expert essaye de deviner la catégorie lexicale des mots inconnus
- Hypothèse utile pour classement propositions
- [P [SN Det Les N* travails] [SV V sont [SA difficiles]]]
  → mot inconnu *travails** est un nom pluriel (masc./fem)

Méthode d'alphacode

- Alphacode : consonnes du mot inconnu dans l'ordre + voyelles, minuscules sans accent: lrstvai
- Restriction: une lettre en moins: rstvai, lstvai etc.
- Elargissement: blrstvai, clrstvai etc.
- 27 requêtes lexique
    - 1 alphacode par mot
    - 0, 1 ou plusieurs résultats par alphacode
- travail (R), travailla (R), travaillai (R), travaillais, travaillas, travaillasse (E), travaillât (R), travaillées (E), travaillés (E), travailles (E), allitératives (E), ravitaillais etc.
- A=6, E=93, R=49, Total=148

Filtrage

- Seulement les propositions qui commencent par la même 1ère lettre
- Conserver propositions les plus proches du mot inconnu: distance lexicographique (nb insertions / effacements / inversions / substitutions pour aller de chaîne A à B)
- Adaptation de l'algoritme de Levenshtein / Damerau
  - Confusion consonne simple / double 10 x moins pénalisée (**imbécilité* ↔ *imbécillité*)
  - Erreurs diacritiques 10 x moins pénalisées (**tres* ↔ *très*)
- Seuil déterminé empiriquement
- Résultat: A=2, E=4, R=4, Total: 10 propositions

Phonétisation

- Prononciation du mot inconnu → recherche phonétique
- Adaptation aux apprenants de L2: substitutions de phonèmes fréquemment confondus
  [ɔ/o], [e/ɛ], [~ɔ/~a], ...
- *Travails** → *travail, travaille, travaillent, travailles*

Méthode ad hoc

- Ebauche de méthode morphologique: substitution du mot entier ou du début / fin de chaîne
- Chevals* → chevaux
- Teniras* → tiendras
- Alleras* → iras
- Fairas* → feras
- Devé* → dû
- Travails* → travaux
- ...

Apostrophe, séparation, casse

- *Sinstaller** → *s'installer*
  Qu', c, d, j, l, m, n, s et *t* peuvent être suivis d'une apostrophe
- Traitement de mots avec apostrophe:
  - *Aujourdhui** → *aujourd'hui, prud'homme, prud'hommal, presqu'île*
- *Unpoisson** → *un poisson*, *weekend* → *week-end*
  Séparation de mots en insérant espace
- Proposition de mettre une majuscule au 1er mot de la phrase

Ordre des propositions

- Calcul d'un score
  - méthode(s)
  - adéquation avec analyse
    - Cat. Lexicale
    - Genre
    - Nombre
- *travaux* = nom pluriel et méthode *ad hoc*

1. travaux
2. travail
3. travailles
4. travaillés
5. travaille
6. travaillas
7. travaillent
8. travailla
9. travaillées
10. travaillai
11. travaillasses
12. travaillai
13. travaillât

Corpus

- 362 soumissions (phrases ou séries), apprenants de Jamaïque, Australie et Canada essentiellement
- Evaluation du corpus, également du point de vue syntaxique

J'	*etais	très	jolie	à	*ecoute	que tu vas venir en Australie!
	DIA		LEX	CPL	DIA LEX	
J'	étais	très	heureuse	d'	*écoute / apprendre*	

Corpus (2)

- 14 494 mots, ~14 mots par phrase
- 2468 erreurs
- 861 mots inconnus

  - 460 propositions correctes du correcteur (53,43%)
  - 213 non-erreurs (24,74%, noms propres, mots inconnus)
  - 188 corrigées manuellement (21,84%, accord, ordre mots, morphologie, lexicales etc.)

Améliorations possibles

- Classement propositions / élimination superflues
- Phonétisation: plusieurs phonétisations alternatives, substitutions
- Elargissement de la méthode ad hoc, méthode morphologique complète.

Conclusions

- Fort potentiel du TAL pour l'apprentissage des langues
- Développement technologique → nombreux logiciels (mobiles) et amélioration du TAL
- Utilisation de techniques robustes (morphologie, étiquettage lexical etc. V. Projets Exills et Mirto)
- Vaincre les réticences et les craintes face à la technologie

Résultats catégories

- Lexicales 916 (37,12%) lexique, nom propre, inconnu, emprunts
- Phonétiques: 643 (26%)
- Accord 454 (18,4%)
- Typographie 390 erreurs (15,8%)
- Mots manquants ou superflus: 338 (13,7%)
- Signe 251 (10,17%) séparation, espace, casse, ponctuation

Résultats méthodes

- Alphacode présent dans 81% des résultats corrects
- Alphacode élargi 15%, restreint: 13%
- Phonétique: 41,5%
- Majuscule: 1,52%
- Ad hoc: 1,09%, séparation 1,09%
- Apostrophe: 0,87%

Analyse et correction grammaticale

- Analyse des phrases d'un texte et représentation (cf FreeText: grammaire en couleurs, arbres syntaxiques)
- Détection d'erreurs: bons résultats pour l'accord, moins fiables dans d'autres domaines
- Prévalence des analyses profondes si on veut une couverture large

Correction ”sémantique”

- Syntaxe parfois insuffisante
  - Constructions équivalentes: Jean semble dormir ↔ il semble que Jean dort. Marie dort-elle? ↔ Est-ce que Marie dort?
  - Constituants manquants
  - Synonymes et antonymes
  - ...
- Possibilité d'utiliser des représentations sémantiques des phrases (plus simple à comparer) (cf MILT (Holland et.al 1993), Athena (Murray 1995), ou comparateur de phrases de cette thèse)