ATALA

Journée d'étude

Le Web comme ressource pour le TAL

N. Gala (DELIC, Univ. Provence)
G. Grefenstette (CEA LIST)

Paris, 11 mars 2006

Le Web comme ressource pour le TAL

"more data are better data" [Church & Mercer 93]

Quelques premiers travaux :

- Acquisition de corpus parallèles [Resnik 98] [Resnik 99]
- Traduction de noms composés [Grefenstette 99]
- Acquisition d'entités nommées [Jacquemin et Bush 00]
- Systèmes question / réponse [Banko, Brill, Dumais et Lin 01]
- Désambiguïsation du rattachement prépositionnel [Volk 01] [Lebarbé 02] [Gala 03]

Le Web comme ressource pour le TAL

Évenements scientifiques récents :

- **Projet WebCorp** (Research and Development Unit for English Studies, University of Central England)
- ***Workshop on Web as a Corpus*,** Corpus Linguistics Conférence, 2005
- ***Workshop on Deep Lexical Acquisition*, ACL-SIGLEX 2005**
- **14e *International World Wide Web Conference* WWW'2005**

Objectifs de la journée

Donner une vision générale des recherches actuelles qui utilisent le Web comme ressource pour différentes tâches liées au traitement automatique des langues :

- Morphologie, syntaxe, sémantique
- Systèmes question / réponse
- Lexicographie et terminologie (monolingue, bilingue)
- Constitution de corpus
- …

Conférencier invité (9h40 – 10h40) :

Adam Kilgarriff

"Bigger and better and bigger and better :

a computational, corpus-driven research programme for linguistics"

Communications du matin (11h – 12h30) :

V. Moriceau, F. Aouladomar (IRIT-CNRS, Toulouse)

"Intérêts d'un corpus issu du Web pour les systèmes Question-Réponse"

B. Grau, I. Robba, A. Vilnat (LIMSI, Orsay)

"Le Web comme source de connaissances pour améliorer la fiabilité des réponses"

T. Lebarbé (LIDILEM, Grenoble)

"Validation des calculs de relations de dépendance : une expérience sur le corpus 'Internet' "

Communications de l'après-midi (14h – 16h30) :

F. Sajous, L. Tanguy (ERSS, Univ. Toulouse)

"Repérage de créations lexicales sur le Web francophone"

S. Léon (DELIC, Univ. Provence)

"Utilisation du Web comme ressource bilingue pour la traduction de termes complexes français/anglais"

L. Deléger (Inserm U729, Paris) , P. Zweigenbaum (Inserm U729, AP-HP, Inalco, Paris)

"Constitution et exploitation d'un corpus parallèle issu du web pour l'extension d'une terminologie multilingue"

T. Delbecque (Inserm U729, Paris), P. Zweigenbaum (Inserm U729, AP-HP, Inalco, Paris)

"Bénéfice d'un catalogue spécialisé de sites internet médicaux pour la constitution de corpus à des fins de recherche"

Posters :

L. Santorum (Univ. Paris 4)

"Rules for the optimisation of the automatic inflexion of Italian "co" and "go" N and Adjs"

P. Saint-Dizier (IRIT-CNRS, Toulouse), S. Zarriess (Univ. Postdam, Allemagne)

"Que peut-on attendre d'un corpus du Web pour caractériser les facettes de l'instrumentalité?"

T. Roy (GREYC, Univ. Caen), P. Beust (GREYC, Univ. Caen)

"Construction et exploration de corpus à partir du Web à l'aide d'une plate-forme logicielle de cartographie documentaire"

C. Fairon (CENTAL, Univ. de Louvain)

"Développement automatisé de corpus spécialisés à partir du Web : l'apport du format RSS"

Adam Kilgarriff

- Traitement automatique des langues, linguistique de corpus et lexicographie.
- Doctorat à l'Université de Sussex autour de la "Polysemie"
- Longman Dictionaries, Oxford University Press, Université de Brighton.
- Actuellement chercheur invité à l'Université de Sussex et directeur de deux entreprises :
  - Lexicography MasterClass Ltd (http://www.lexmasterclass.com)
  - Lexical Computing Ltd (http://www.sketchengine.co.uk)

Adam Kilgarriff

- Initiateur du projet SENSEVAL (automatic word sense disambiguation),
- Membre d'EURALEX (European Association for Lexicography)
- Ancien president d'ACL-SIGLEX (Association for Computational Linguistics Special Interest Group on the Lexicon).
- **Intérêt pour rendre le Web accessible comme corpus pour les linguistes.**

Adam Kilgarriff

"Bigger and better and bigger and better :
a computational, corpus-driven research programme for linguistics"

Quelques références (1)

Banko, M., Brill, E., Dumais, S. et Lin, Ng. (2001) *Data Intensive Question Answering*. TREC-10 Notebook, Gaithersburg, USA.

Church, K. W. et Mercer, R. (1993) *Introduction to the special issue on computational linguistics using large corpora*. Computational Linguistics 19(1):1-24.

Gala, N. (2003) Un modèle d'analyseur syhtaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires. Thèse de doctorat, Université Paris XI.

Grefenstette, G. (1999) *The World Wide Web as a resource for example-based machine translation tasks*. Dans Proc. of Aslib Conference on Translation and the Computer.

Jacquemin, C. et Bush, C. (2000) *Combining lexical and formatting cues for named entity acquisition from the Web*. Dans Proc. of joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), édité par H. Schutze, Hong Kong.

ATALA

Quelques références (2)

Lebarbé, T. (2002) Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative. Thèse de doctorat, Université de Caen.

Resnik, P. (1998) *Parallel strands : a preliminary investigation into mining the Web for bilingual text*. Dans Proc. of the third Conference of the Association for Machine Translation in the Americas, AMTA-98, in lecture notes in Artificial Intelligence, 1529, Langhorne, PA, pp. 28-31.

Resnik, P. (1999) *Mining the Web for bilingual text*. Dans Proc. of the 37th Annual Meeting of the ACL, June.

Volk, M. (2001) *Exploiting the Web as a corpus to resolve preprositionnal attachment*. Dans Proc. of Conference on Corpus Linguistics, Lancaster, pp. 601-606.