COLING 2004

The 20th International Conference on Computational Linguistics
Post-Conference Workshop

Proceeding of the Workshop on Multilingual Linguistic Ressources MLR2004

Editors:
Gilles Sérasset, Susan Armstrong, Christian Boitet,
Andrei Popescu-Belis, Dan Tufis

August 28th, 2004
University of Geneva, Switzerland

Foreword

In an ever expanding information society, most information systems are now facing the "multilingual challenge". Multilingual language resources play an essential role in modern information systems. Such resources need to provide information on many languages in a common framework and should be (re)usable in many applications (for automatic or human use).

Many centres have been involved in national and international projects dedicated to building harmonised language resources and creating expertise in the maintenance and further development of standardised linguistic data. These resources include dictionaries, lexicons, thesauri, word-nets, and annotated corpora developed along the lines of best practices and recommendations. However, since the late 90's, most efforts in scaling up these resources remain the responsibility of the local authorities, usually, with very low funding (if any) and few opportunities for academic recognition of this work. Hence, it is not surprising that many of the resource holders and developers have become reluctant to give free access to the latest versions of their resources, and their actual status is therefore currently rather unclear.

The goal of this workshop is to study problems involved in the development, management and reuse of lexical resources in a multilingual context. Moreover, this workshop provides a forum for reviewing the present state of language resources. The workshop is meant to bring to the international community qualitative and quantitative information about the most recent developments in the area of linguistic resources and their use in applications.

The impressive number of submissions (38) to this workshop and in other workshops and conferences dedicated to similar topics proves that dealing with multilingual linguistic ressources has become a very hot problem in the Natural Language Processing community.

To cope with the number of submissions, the workshop organising committee decided to accept 16 papers from 10 countries based on the reviewers' recommendations. Six of these papers will be presented in a poster session. The papers constitute a representative selection of current trends in research on Multilingual Language Resources, such as multilingual aligned corpora, bilingual and multilingual lexicons, and multilingual speech resources. The papers also represent a characteristic set of approaches to the development of multilingual language resources, such as automatic extraction of information from corpora, combination and re-use of existing resources, online collaborative development of multilingual lexicons, and use of the Web as a multilingual language resource.

The development and management of multilingual language resources is a long-term activity in which collaboration among researchers is essential. We hope that this workshop will gather many researchers involved in such developments and will give them the opportunity to discuss, exchange, compare their approaches and strengthen their collaborations in the field.

The organisation of this workshop would have been impossible without the hard work of the program committee who managed to provide accurate reviews on time, on a rather tight schedule. We would also like to thank the Coling 2004 organising committee that made this workshop possible. Finally, we hope that this workshop will yield fruitful results for all participants.

Gilles Sérasset
Organising chair
GETA (Study Group for Machine Translation), CLIPS-IMAG laboratory
Université Joseph Fourier, France

Program Committee

Gilles Sérasset *(Chair)*	GETA CLIPS-IMAG, Université Joseph Fourier - Grenoble I, France
Susan Armstrong	ISSCO, Université de Genève, Switzerland
Pushpak Battacharya	IIT, Mumbai, India
Igor Boguslavski	IITP, Moscow, Russia
Christian Boitet	GETA CLIPS-IMAG, Université Joseph Fourier - Grenoble I, France
Pierrette Bouillon	ISSCO, Université de Genève, Switzerland
Jim Breen	Monash University, Australia
Nicoletta Calzolari	CNR, Pisa, Italy
Dan Cristea	University Al.I.Cuza Iasi, Romania
Patrick Drouin	OLST, University of Montreal,Canada
Sanae Fujita	NTT, Kyoto, Japan
Ulrich Heid	IMS-CL, University of Stuttgart, Germany
Hitoshi Isahara	CRL, Nara, Japan
Kyo Kageura	NII, Tokyo, Japan
Chuah Choy Kim	USM, Penang, Malaisie
Mathieu Mangeot	NII, Tokyo, Japan
Alain Polguère	OLST, University of Montreal,Canada
Andrei Popescu-belis	ISSCO, Université de Genève, Switzerland
Jean Senellart	SYSTRAN, France
Mandel Shi	Xiamen University, China
Virach Sornlertlamvanich	Thai Computational Linguistics Laboratory, CRL, Thailand
Pr. Kumiko Tanaka-Ishii	Tokyo University, Japan
Philippe Thoiron	CRTT, Université de Lyon 2, France
Dan Tufis	RACAI, Uni Bucharest, Romania
Michael Zock	LIMSI, Orsay, France

Tentative Program

Time	Event
08:30 — 09:00	Registration & Welcome
09:00 — 10:00	Paper Session • **JMdict: a Japanese-Multilingual Dictionary** • **A Generic Collaborative Platform for Multilingual Lexical Databases Development**
10:00 — 11:00	Poster Session and opened discussions • **Semi-Automatic Construction of Korean-Chinese Verb Patterns based on Translation Equivalency** • **Bilingual Sign Language Dictionary to Learn the Second Sign Language without Learning a Target Spoken Language** • **Building Parallel Corpora for eContent Professionals** • **Revising the WordNet Domains Hierarchy: semantics, coverage and balancing** • **PolyphraZ: a tool for the management of parallel corpora** • **Multilingual Text Induced Spelling Correction**
11:00 — 11:30	Coffee Break
11:30 — 13:00	Papers Session • **A Model for Fine-Grained Alignment of Multilingual Texts** • **Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora** • **Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications**
13:00 — 14:00	Lunch Break

Time	**Event**
14:00 — 15:00	Papers Session • **A Method of Creating New Bilingual Valency Entries using Alternations** • **Automatic Construction of a Transfer Dictionary Considering Directionality**
15:00 — 15:30	Poster Session and opened discussions • **Semi-Automatic Construction of Korean-Chinese Verb Patterns based on Translation Equivalency** • **Bilingual Sign Language Dictionary to Learn the Second Sign Language without Learning a Target Spoken Language** • **Building Parallel Corpora for eContent Professionals** • **Revising the WordNet Domains Hierarchy: semantics, coverage and balancing** • **PolyphraZ: a tool for the management of parallel corpora** • **Multilingual Text Induced Spelling Correction**
15:30 — 16:00	Coffee Break
16:00 — 17:30	Papers Session • **Building and sharing multilingual speech resources, using ERIM generic platforms** • **Multilinguality in ETAP-3: Reuse of Lexical Resources** • **Qualitative Evaluation of Automatically Calculated Acception Based MLDB**
17:30 — 18:00	Opened discussions & Closing

Contents

Authors Index