AEPC 2011

Proceedings of
The Second Workshop on Annotation and Exploitation of Parallel Corpora

associated with

The 8th International Conference on
Recent Advances in Natural Language Processing
(RANLP 2011)

15 September, 2011
Hissar, Bulgaria

Preface

This workshop is a follow-up of the First Workshop on Annotation and Exploitation of Parallel Corpora (http://math.ut.ee/tlt9/aepc/).

The creation of parallel corpora has been very active especially since 90s. The globalization, the extension of EU with new countries as well as the availability of open-source places for information, such as Wikipedia, DBPeadia, etc. required a multilingual approach towards the interpersonal and official communication. This status quo produced a lot of parallel data – especially administrative and political documents in several languages (EuroParl), but also news (SETIMES) and texts on various topics (wikipedia, bi- and multilingual web sites). However, the fast compilation of large amounts of data very often compromised in lower quality of paralleling texts. Here comes the challenge to discover the inconsistencies in these huge quantities of parallel data, to process them in adequate ways, and to exploit them for various applications: QA, Information Retrieval, Machine Translation, etc. The parallel corpora go beyond word-to-word alignments. They rely on dependency, constituent or semantic pairings. There appeared guidelines and tools for aligning linguistic structures, which raised the issue of transferability of aligning schemes from one language to another, and also for the compatibility among various resources.

The topics, which fall within the scope of the workshop, include: Strategies for creation of annotated parallel corpora; Annotation guidelines for alignment; Annotation alignment transfer over languages; Tools for manual and automatic processing and exploitation of parallel corpora; Problems in manual and automatic alignment; Syntax-based and semantic-based approaches to using parallel corpora in MT; Parallel Grammars; Parallel Statistical Parsing; Usability of the existing parallel resources for various applications.

The workshop has been supported by the European project EuroMatrixPlus – Bringing Machine Translation for European Languages to the User.

The Organizers

Organizers:

Kiril Simov (IICT, Bulgarian Academy of Sciences)
Petya Osenova (Sofia University “St. Kl. Ohridski” and IICT Bulgarian Academy of Sciences)
Radovan Garabik (JÚL’Š, Slovak Academy of Sciences)
Jürg Tiedemann (Uppsala University)

Program Committee:

António Branco (University of Lisbon)
Nicoletta Calzolari (Institute of Computational Linguistics of the National Research Council)
Koenraad De Smedt (University of Bergen)
Dan Flickinger (Stanford University)
Dale Gerdemann (University of Tübingen)
Voula Giouli (Institute for Language and Speech Processing)
Silvia Hansen (University of Mainz)
Erhard Hinrichs (University of Tübingen)
Valia Kordoni (University of Saarland)
Vladislav Kubon (Charles University)
Lothar Lemnitzer (Berlin-Brandenburg Academy of Sciences and Humanities)
Preslav Nakov (National University of Singapore)
Cristina Vertan (University of Hamburg)
Eline Westerhout (University of Utrecht)

Invited Speaker:

Preslav Nakov (National University of Singapore)

Table of Contents

Reusing Parallel Corpora between Related Languages
Preslav Nakov 1

Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying
Marilisa Amoia, Kerstin Kunz and Ekaterina Lapshinova-Koltunski 2

A tagged and aligned corpus for the study of Proper Names in translation
Emeline Lecuit, Denis Maurel and Duško Vitas 11

Building the multilingual TUT parallel treebank
Manuela Sanguinetti and Cristina Bosco 19

Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment
Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva 29

Parallel Corpora in Aspectual Studies of Non-Aspect Languages
Maria Stambolieva 39

Coreference Annotator - A new annotation tool for aligned bilingual corpora
Mara Tsoumari and Georgios Petasis 43

Using Manual and Parallel Aligned Corpora for Machine Translation Services within an On-line Content Management System
Cristina Vertan and Monica Gavrila 53

Workshop Program

Friday, 16 September 2011

9:20–9:30	Opening
9:30–10:30	*Reusing Parallel Corpora between Related Languages* Preslav Nakov
10:30–11:00	Coffee Break
11:00–11:30	*Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying* Marilisa Amoia, Kerstin Kunz and Ekaterina Lapshinova-Koltunski
11:30–12:00	*Parallel Corpora in Aspectual Studies of Non-Aspect Languages* Maria Stambolieva
12:00–12:30	*Coreference Annotator - A new annotation tool for aligned bilingual corpora* Mara Tsoumari and Georgios Petasis
12:30–14:00	Lunch
14:00–14:30	*A tagged and aligned corpus for the study of Proper Names in translation* Emeline Lecuit, Denis Maurel and Duško Vitas
14:30–15:00	*Using Manual and Parallel Aligned Corpora for Machine Translation Services within an On-line Content Management System* Cristina Vertan and Monica Gavrila
15:00–15:30	*Building the multilingual TUT parallel treebank* Manuela Sanguinetti and Cristina Bosco
15:30–16:00	Coffee Break
16:00–16:30	*Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment* Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva
16:30–16:40	Closing Remarks