INVITED SPEAKER:

Elliot Macklovitch, University of Montreal

ORGANIZING COMMITTEE:

Rada Mihalcea, University of North Texas
Ted Pedersen, University of Minnesota, Duluth

PROGRAM COMMITTEE:

Lars Ahrenberg, Linköping University
Nicoletta Calzolari, University of Pisa
Tim Chklovski, Massachusetts Institute of Technology
Mona Diab, University of Maryland
Ulrich Germann, USC Information Sciences Institute
Daniel Gildea, University of Pennsylvania
Maria das Graças Volpe Nunes, University of São Paulo
Nancy Ide, Vassar College
Philippe Langlais, University of Montreal
Lucia Helena Machado Rino, Federal University of São Carlos
Eduard Hovy, USC Information Sciences Institute
Elliot Macklovitch, University of Montreal
Daniel Marcu, USC Information Sciences Institute
Dan Melamed, New York University
Magnus Merkel, Linköping University
Ruslan Mitkov, University of Wolverhampton
Hermann Ney, RWTH Aachen
Grace Ngai, Hong Kong Polytechnic University
Franz Och, USC Information Sciences Institute
Kemal Oflazer, Sabanci University
Kishore Papineni, IBM
Jessie Pinkham, Microsoft Research
Andrei Popescu-Belis, ISSCO/TIM/ETI University of Geneva
Florence Reeder, MITRE
Philip Resnik, University of Maryland
Antonio Ribeiro, European Commission, Joint Research Centre, Italy
Michel Simard, University of Montreal
Harold Somers, University of Manchester Institute of Science and Technology
Arturo Trujillo, Canon Research Centre Europe
Dan Tufiş, RACAI, Romania
Jean Véronis, University of Provence
Clare Voss, Army Research Lab
Yorick Wilks, University of Sheffield

WORKSHOP WEBSITE:

http://www.cs.unt.edu/~rada/wpt/

INTRODUCTION

This volume contains papers accepted for presentation at the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond. This event took place on May 31, 2003, in Edmonton, Canada, immediately following the HLT-NAACL Human Language Technology Conference.

The workshop was centered around the problem of building and using parallel corpora, which are vital resources for efficiently deriving multi-lingual text processing tools. We issued calls for both regular and short, late–breaking papers. After careful review by our program committee, nine regular papers and three short papers were accepted for presentation. We were truly impressed by the high quality of the reviews provided by all the members of the program committee, particularly since deadlines were very tight. All of the committee members provided timely and thoughtful reviews, and the papers that appear have certainly benefited from that expert feedback.

Before the workshop, we conducted a comparative evaluation of word alignment techniques that was open to anyone with a word alignment system and a bit of courage. English–French and Romanian–English parallel text was made available to participants, and they had about two weeks to develop or train their systems on this data. Then previously held out test data was released, and participants had one week to submit word aligned versions of the test data to us for scoring. Seven teams participated in the shared task, and each has a fully reviewed short paper in the proceedings describing their system. In addition, there are two papers that present overall evaluations of the exercise.

The shared task on word alignment would not have been possible without parallel text, and there are quite a few people who deserve thanks for creating data that was a part of this task.

The English–French parallel text came to us from Franz Och, Herman Ney, and Ulrich German. Franz and Herman kindly provided their word aligned English–French corpus, which served as our gold standard in the English–French evaluation. The training data consisted of Ulrich Germann's freely available Aligned Hansards of the 36th Parliament of Canada (Release 2001-1a).

The Romanian–English gold standard data was created by a number of student volunteers from the Department of English, Babes-Bolyai University, Cluj-Napoca, Romania. The training data includes the freely available MULTEXT-East Romanian–English version of George Orwell's novel 1984, as well as the Romanian Constitution and newspaper articles collected from the Internet.

Even with all that data, there would not have been a shared a task without participants. We give our deepest thanks to the shared task participants, who took a leap of faith in allowing their systems to be part of a comparative evaluation.

Finally, when we first started planning this workshop, we agreed that having a high quality invited speaker was crucial. We nearly immediately realized that there would be no better choice than Elliot Macklovitch, who has a wealth of experience working with parallel text. We were pleasantly surprised when Elliot quickly and enthusiastically agreed to deliver the invited talk. We thank him not only for his talk, but also for the boost of confidence his quick acceptance of our invitation provided.

Ted Pedersen
Rada Mihalcea
April 2003