SPONSORS:

Advanced Research and Development Activity (ARDA)
Advanced Question Answering for Intelligence (AQUAINT) program
Defense Advanced Research Projects Agency (DARPA)
Translingual Information Detection, Extraction and Summarization (TIDES) program

INVITED SPEAKERS:

James Cowie, Director, Computing Research Laboratory, New Mexico State University
Randall Flynn, Geographer, National Imagery and Mapping Agency, and
Executive Secretary for Foreign Names, US Board on Geographic Names

PROGRAM COMMITTEE:

András Kornai, Metacarta, Co-chair
Beth Sundheim, SPAWAR Systems Center (US Dept. of Navy), Co-chair

Doug Appelt, SRI International
Merrick Lex Berman, Harvard Yenching Institute
Sean Boisen, BBN Technologies
Quintin Congdon, US Army National Ground Intelligence Center
James Cowie, New Mexico State University Computing Research Laboratory
Linda Hill, University of California, Santa Barbara
Douglas Jones, MIT Lincoln Laboratory
George Wilson, MITRE

CONFERENCE WEBSITE:

http://www.kornai.com/NAACL

PREFACE

The analysis of geographic references in natural language text involves, at least conceptually, four distinct stages. Of course, implementations may vary greatly in how these stages are interleaved. The first conceptual stage is *geographic entity reference detection*: strings such as *New York, the Amazon delta, LaGuardia, the San Diego-Tijuana border, [the] Brooklyn Bridge, a mile from downtown Manhattan*, etc. are identified in the text (Rauch et al). Second, *contextual information gathering* may help identify the type and approximate location of geographic entities: LaGuardia *Airport* vs. LaGuardia *Community College*, the *town of* Manhattan *(population 44,831)*, etc. (Manov et al, Bilhaut et al). Third is the actual *disambiguation* of the entity with respect to both type (*New York City* vs. *New York State*) and location (*Orange County, California* vs. *Orange County, Florida*) (Leidner et al, Waldinger et al, Li et al).

Up to this point we can proceed, at least in principle, entirely on the basis of linguistic reasoning about the document at hand, but the fourth stage, *grounding*, which is the assignment of geographic coordinates to the identified entities, requires background knowledge in some form or another. The single most important knowledge resource is a *gazetteer* containing at least longitude and latitude data associated with each placename, and possibly supplementary information such as elevation, population, province (or state), type of location, variant spellings, etc. There are large electronic gazetteers in existence that are publicly available – one of the workshop goals was to discuss how they can be tailored and exploited to meet the needs of NLP (Kwok and Deng, Axelrod) and conversely, how NLP techniques can be used for *database population*, the addition of (partial) information about hitherto missing entries to the gazetteer (Uryupina). Other background information, such as a collection of grounded documents, may also prove useful (Smith and Mann, Li et al) both for training and evaluation purposes.

In organizing the workshop into two broad sessions we did not follow the above stages rigidly, especially as these stages are most relevant for analyzing run-time behavior, while most papers deemphasized run-time processing detail (testing) and concentrated on creating and maintaining the background knowledge they require (training). Our first session gathered those papers that focused on the semantics of geographic references (feature types, ontologies, disambiguation), and the second session included those that emphasized systems and gazetteer development. An invited talk kicked off each of the two sessions, and a discussion period ended each one.

Effective analysis of textual references to places is a critical core technology for a wide variety of NLP applications. As the 12 papers selected by the Program Committee for presentation in this volume (from a total of 19 submissions) demonstrate, this is a very active research area, and one that feels the very same tensions as the rest of NLP. There seems to be kind of Boyle-Mariotte Law (volume times pressure is constant) in operation: at one end of the scale, researchers apply a lot of "pressure" (deep analysis) to a few dozen to few thousand items; at the other end, they apply analytic techniques that are a great deal shallower, but the "volume" of items is considerably larger, often in the millions to hundreds of millions of items.

To a large extent, the two ends of the scale correspond to the well-articulated *rationalist* and *empiricist* approaches to NLP, but here this seems to be dictated more by the nature of the data at hand than by overriding philosophical considerations. In this collection, the high pressure/low volume extreme is exemplified by papers such as those of Bilhaut et al, Waldinger et al, Smith and Mann, Manov et al, and Southall. The low pressure/high volume extreme is seen in Li et al and Rauch et al, with papers like those of Leidner et al and Kwok and Deng falling somewhere in the middle.

The workshop brought together researchers from the NLP and Digital Library communities; missing were researchers from the GIS community. Map-based visualization of the results of the grounding state of analysis, including the capability for bi-directional (both text-based and map-based) querying, while obviously of prime concern to the designers of several systems described here, is discussed in detail only in one of the twelve papers (Leidner et al). Perhaps the call for papers was too restrictive in this direction, and there is an opportunity for a more user-centric follow-on workshop. Yet as organizers we feel that advances in core technology that can bring to bear both lexical and spatial background knowledge are more critical than advances in visualization, if we are to respond to the challenges of providing accurate analyses of geographic references in broad domains and across languages, and to provide useful information on subjects for which there is sparse training data.

To support analysis in multilingual and cross-lingual settings, such advances must of necessity begin with the most mundane aspects of the problem: when to bracket a text string (tagging guidelines), how to deal with foreign names in native script or in transliteration (text normalization), etc. Recognition of the various ways that a given place may be referenced in one language is a challenging problem in

and of itself, and issues of name translation and transliteration and special character sets multiply that problem. While our workshop revealed significant progress (Southall in identifying major relation types for tagging, Axelrod from the database infrastructure standpoint, and Kwok and Deng concerning the important special case that is Chinese), clearly much more remains to be done.

Work on applications such as question-answering, multidocument summarization and information extraction, or first-story detection in streams of broadcast news, requires solid semantic foundations, and again the workshop has some progress to report (Manov et al on ontology development, Waldinger et al on axiomatic deduction, Smith and Mann on type classification), but here, perhaps, even more remains to be done.

We would like to thank the authors, the members of the Program Committee, and the invited speakers for their contributions to the planning and execution of the workshop, the workshop sponsors for their financial support, and the HLT/NAACL conference organizers, especially Ed Hovy, James Allen, Steven Abney, and Dragomir Radev, for their significant contributions to the overall management of the workshop and their direction in preparing the publication of the proceedings.

András Kornai and Beth Sundheim
May 2003