Resources for the Millennium
Panel Session at LREC2000

Catherine Macleod

Computer Science Department
New York University, 715 Broadway, 7th Floor
New York, New York 10003-6806, USA
macleod@cs.nyu.edu

Abstract

This panel will discuss what they see as the essential resources needed to support Natural Language Processing research in the future. This is an opportunity to explore questions of where NLP research is heading and to plan for the creation of appropriate resources. Looking ahead is necessary for builders of resources since they are costly in terms of both time and money. This community should concentrate their efforts on creating resources that researchers will need. This panel is a modest attempt to identify some of these needs and to inspire us to find ways meet them.

1. Purpose of the Panel

Resources are costly and time-consuming to build. Often resource makers are chided about building resources without consulting others in the scientific community. We think that there should be a much closer connection between those who create resources and those who use them. We are bringing together NLP researchers from the fields of speech recognition, machine translation and information extraction to discuss the resources that they believe will be needed in the next millennium.

2. Some Topics to be Discussed

2.1. Text Resources

We will discuss the GSK (the Japanese Linguistic Resource Consortium), the Kyoto Corpus which is an analyzed corpus based on Japanese newspaper articles and the new Japanese governmental MT project which will make a corpus of parallel and comparable texts. The difficulties of assembling such a corpus should be noted.

2.2. Speech Resources

One of the most important issues for speech recognition is how to create language models (rules) for spontaneous speech. When recognizing spontaneous speech, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, and repetitions. In order to build language models for spontaneous speech, we need to have large spontaneous speech corpora. In Japan, a Science and Technology Agency Priority Program (Organized Research Combination System) entitled ”Spontaneous Speech: Corpus and Processing Technology” is being developed and the aims of this project will be discussed. We will also be touching on the question of spoken dialogue resources and the issues related to their creation.

2.3. Tools

A much neglected topic in the discussion of resources are the tools which facilitate corpus collection and annotation. We need – in addition to archival corpora – to become much more agile in corpus creation and annotation. Since data is the life-blood of modern computational linguistics, we need to be able to decentralize data collection and enable people to rapidly collect and annotate (and SHARE) the data that they need. The way to do this is by easy-to-use public domain tools (e.g., MATE, Alembic Workbench,...) and by shared STANDARDS. This would allow people to quickly collect data from various on-line information feeds, do hand-corrected automatic annotation, and then use shared (standardized) evaluation methods to evaluate performance.

3. Panel Participants

Chair:

Catherine Macleod
New York University

Participants:

Sadaoki Furui
Tokyo Institute of Technology

Lynette Hirschman
Mitre

Sadao Kurohashi
Kyoto University

Masumi Narita
RICOH

Antoine Ogonowski
Lexiquest

Roberto Pieraccini
Speechworks

Marilyn Walker
AT&T Research