TIPSTER TEXT PROGRAM

PHASE II

Proceedings of a Workshop held at

Vienna, Virginia

May 6-8, 1996

Sponsored by:

Defense Advanced Research Projects Agency

This document contains copies of reports prepared for the DARPA TIPSTER Text Program - Phase II Workshop. Included are reports from DARPA sponsored program and other materials prepared for use at the workshop.

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

Table of Contents

Section A - Overview of the TIPSTER Program

Welcome to TIPSTER 1

Focus of TIPSTER Phases I and II 9
Patricia A. Currier, Patrick J. Altomari

Evaluation Driven Research: The Foundation of the TIPSTER Text Program 13
Dr. John D. Prange

Technology Transfer: Some Observations from the TIPSTER Text Program 23
Dr. Sarah M. Taylor

Architecture Overview 33
TIPSTER SE/CM

The Message Understanding Conferences 35
Beth M. Sundheim (NCCOSC)

The Text Retrieval Conferences (TRECs) 39
Donna Harman

TIPSTER Phase III Goals 41
F. Ruth Gee (ORD)

Section B - TIPSTER Phase II Contract Summaries

TIPSTER II Activities at HNC 45
Marc R. Ilgen, David A. Rushall (HNC Software, Inc.)

The Lockheed Martin TIPSTER II Project 47
Steven Maiorano (ORD)

Table of Contents

CERVANTES - A System Supporting Text Analysis 49
Jim Cowie (CRL, NMSU)

The NYU TIPSTER II Project 51
Dr. Sarah M. Taylor

SRA Participation in TIPSTER Phase II 53
Lisa Rau

The SRI TIPSTER II Project 55
Steven Maiorano (ORD)

SPOT: TRW's Multi-Lingual Text Search Tool 57
Peggy Otsubo (TRW)

Section C - Projects Employing TIPSTER Technologies

Advanced Data Extraction and Preparation Via TIPSTER (ADEPT) 61
John Kielty, Ira Sider (LM)

Cable Abstracting and INdexIng System (CANIS) Prototype 69
Ira Sider, Jeffrey Baker, Deborah Brady, Lynne Higbie, Tom Howard (LM)

Management of Free Text for NDIC: An Overview of the
FTM Project 77
Nancy J. Pruett, Thomas Kinsella (ORD, NDIC)

The HOOKAH Information Extraction System 79
Chris Barclay, Sean Boisen, Clinton Hyde, Ralph Weischedel (BBN)

Message Handler 83
Dr. Steven Flank (DARPA)

OLEADA: User-Centered TIPSTER Technology for
Language Instruction 85
William C. Ogden, Philip Bernick (CRL, NMSU)

Table of Contents

An Overview of the Prototype Information Dissemination System (PRIDES) 91
Virginia Cevasco (Logicon)

SPOT: TRW's Multi-Lingual Text Search Tool 95
Peggy Otsubo (TRW)

The Temple Translator's Workstation Project 101
Rémi Zajac (CRL, NMSU); Michelle Vanni (DoD)

SRA Project for ARPA/USACOM 107
David Oppenheimer (SRA)

Chinese Information Extraction & Retrieval 109
Sean Boisen, Michael Crystal, Erik Peterson, Ralph Weischedel (BBN); John Broglio, Jamie Callan, Bruce Croft (Umass); Theresa Hand, Thomas Keenan, Mary Ellen Okurowski (DoD)

TIPSTER-Compatible Projects at Sheffield 121
Hamish Cunningham, Kevin Humphreys, Robert Gaizauskas, Yorick Wilks

Section D - Research in Information Extraction & Document Detection

Progress in Information Extraction 127
Ralph Weichedel, Sean Boisen, Daniel Bikel, Robert Bobrow, Michael Crystal, William Ferguson, Allan Wechsler, and the PLUM Research Group

The Role of Syntax in Information Extraction 139
Ralph Grishman (NYU)

Natural Language Information Retrieval: TIPSTER-2 Final Report 143
Tomek Strzalkowski (GE)

Recent Advances in HNC's Context Vector Information Retrieval Technology 149
Marc R. Ilgen, David A. Rushall

Table of Contents

A Context Vector-Based Self Organizing Map for Information Visualization 159
David A. Rushall, Marc R. Ilgen

A Simple Probabilistic Approach to Classification and Routing 167
Louise Guthrie, James Leistensnider (LM)

An Evaluation of Co-Reference Resolution Strategies for Acquiring Associated Information 179
Lois C. Childs (LM)

Advances in Multi-Lingual Text Retrieval 185
Mark Davis (CRL, NMSU)

Integration of Document Detection and Information Extraction 195
Louise Guthrie (LM), Tomek Strzalkowski, Fang Lin, Wang Jin (GE)

SRI's TIPSTER II Project 201
Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Andrew Kehler, Mark Stickel, Mabry Tyson (SRI)

Section E - TIPSTER Architecture

Overview and Accomplishments: The SE/CM Perspective 211
TIPSTER SE/CM

Building an Architecture: A CAWG Saga 213
Ralph Grishman (NYU)

The Architecture Demonstration System 217
TIPSTER SE/CM

TUIT: A Toolkit for Constructing Multi-Lingual TIPSTER User Interfaces . . 219
William C. Ogden (CRL, NMSU)

TIPSTER Text Phase II Architecture Concept 221
Architecture Committee

Table of Contents

TIPSTER Text Phase II Architecture Design 249
Bill Caid, Jamie Callan, Jim Conley, Harold Corbin, Jim Cowie, Kathy DiBella, Ted Dunning, Joe Dzikiewicz, Louise Guthrie, Jerry Hobbs, Clint Hyde, Mark Ilgen, Paul Jacobs, Matt Mettler, Bill Ogden, Peggy Otsubo, Bev Schwartz, Ira Sider, Ralph Weischedel, Remi Zajac (Architecture Committee)

TIPSTER Text Phase II Architecture Requirements 307
Architecture Committee

Configuration Management Plan 345
Architecture Committee

Section F - Evaluating the Technologies: The Text REtrieval Conferences (TREC)

Summary Results of TREC-3 and TREC-4 373
Donna Harman

Section G - Evaluating the Technologies: The Message Understanding Conferences (MUC)

Design of the MUC-6 Evaluation 413
Ralph Grishman (NYU), Beth M. Sundheim (NCCOSC)

Overview of Results of the MUC-6 Evaluation 423
Beth M. Sundheim (NCCOSC)

Section H - Multi-Lingual Entity Task

The Multi-Lingual Entity Task (MET) Overview 445
Roberta Merchant, Mary Ellen Okurowski (DoD); Nancy Chinchor (SAIC)

Multi-Lingual Entity Task (MET): Japanese Results 449
Steven Maiorano (ORD), Terry Wilson (DoD)

Table of Contents

An Interpretative Data Analysis of Chinese Named Entity Subtypes 453
Thomas A. Keenan (DoD)

The Multi-Lingual Entity Task: A Descriptive Analysis of ENAMEX in Spanish 457
Don D. Anderson (DoD)

MITRE: Description of the ALEMBIC System as Used in MET 461
John Aberdeen, John Burger, David Day, Lynette Hirschman, David Palmer, Patricia Robinson, Marc Vilain (MITRE)

NameTag™ Japanese and Spanish Systems as Used for MET 463
Chinatsu Aone (SRA)

Approaches in MET (Multi-Lingual Entity Task) 465
Damaris Ayuso, Daniel Bikel, Tasha Hall, Erik Peterson, Ralph Weischedel (BBN); Patrick Jost (FinCEN)

CRL's Approach to MET 467
Jim Cowie (CRL, NMSU)

NTT Data: Description of the Erie System Used for MUC-6 469
Yoshio Eriguchi, Tsuyoshi Kitani (NTT Data)

MET Name Recognition with Japanese FASTUS 471
Megumi Kameyama (SRI)

Description of NEC/Sheffield System Used for MET Japanese 475
Yoshikazu Takemoto, Takahiro Wakao, Hiroshi Yamada, Robert Gaizauskas, Yorick Wilks

Section I - TIPSTER References

How to Get Information about TIPSTER 479

List of Available Documents 481