Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications

ACL/EACL-97 Workshop Proceedings
July 12th 1997

Madrid

Editors
Piek Vossen (Chair)
Geert Adriaens
Nicoletta Calzolari
Antonio Sanfilippo
Yorick Wilks

Organized under the auspices of the Language Engineering section of the
European Commission, Directorale General XIII Luxembourg,
by the projects EuroWordNet(LE2 4003), Sparkle (LE1 2111) and Ecran

Order additional copies from:

ACL
P.O. Box 6090
Somerset, NJ, 08875 USA
+1-908-873-3898
acl@bellcore.com

Preface

In the past years the development of high-quality and overall language resources has been the focus of many research groups. More recently also the corpus-based extraction of such resources has gained a wider interest. EuroWordNet, Sparkle and Ecran try to package some of this know-how and expertise into state-of-the-art tools and resources that can directly be applied in NLP-based services. In the EuroWordNet project a multilingual database is developed with wordnets for four European Languages linked to the existing Princeton WordNet (version 1.5). Such a database can be used in multilingual retrieval applications but it can also be seen as a starting point for automatic-translation aids, inferencing systems, and information extraction systems. Sparkle and Ecran both address the creation of language resources and technologies for real-world NLP applications in parallel. This objective is carried out through the development of software tools in the areas of shallow parsing and lexical acquisition. These tools are used to induce linguistic knowledge from text corpora and are progressively enriched by the information acquired.

In all three projects the current limits of Linguistic Technology are being explored for their practical benefits. Whereas EuroWordNet aims at the broadening and extension of the Princeton WordNet to a generic multilingual resource which is the first in its kind, Sparkle and Ecran aim at the dynamic anchoring of resources and information to the data and corpora that are of a user's interest. The availability of these resources and tools is essential for the new generation of applications and products dealing with information in electronic form. The projects have finished their specification phase and are in the process of generating the results. In this workshop we want to discuss the scope and formats of semantic resources and information acquisition tools with scholars in the field and researchers from commercial R&D departments who have experience in developing and using them. The main themes of the workshop are:

- compatibility and standards of multilingual semantic resources and lexical acquisition tools.
- the validation of multilingual semantic resources and lexical acquisition tools.
- performances of semantic resources and lexical acquisition tools in NLP tasks.
- partial or phrasal parsing of text.
- linking text with lexical databases: sense-differentiation,
- sense-tagging and sense-disambiguation tasks, domain-differentiation of text and lexical resources.

The first three papers in the proceedings address issues related to the building and checking of lexical semantic resources. The remainder of the papers mainly deal with the application of lexical semantic resources in various NLP tasks, ranging from information retrieval, semantic tagging and information extraction, or they deal with the extraction of information from text-corpora to build such resources eventually.

ORGANIZING COMMITTEE

Piek Vossen,
Computer Centrum Letteren, University of Amsterdam
e-mail: Piek.Vossen@let.uva.nl
Cintha Harjadi,
Computer Centrum Letteren, University of Amsterdam
e-mail: Cintha.Harjadi@let.uva.nl
Horacio Rodriquez,
Universitat Politecnica de Catalunya,
e-mail: Horacio@lsi.upc.es

PROGAM COMMITTEE

Piek Vossen,
University of Amsterdam, The Netherlands,
e-mail: Piek.Vossen@let.uva.nl
Nicoletta Calzolari,
Istituto di Linguistica Computazionale del CNR, Italy,
e-mail: glottolo@vm.cnuce.cnr.it
Antonio Sanfilippo,
Sharp Laboratories, UK,
email: Antonio.Sanfilippo@sharp.co.uk
Geert Adriaens,
Novell Linguistic Development, Belgium,
e-mail: Geert_Adriaens@novell.com
Yorick Wilks,
University of Sheffield, UK,
e-mail: yorick@dcs.shef.ac.uk

CONTENTS