

## Topics and Motivation

SEMANTIC ANNOTATION is augmentation of data to facilitate automatic recognition of the underlying semantic structure. A common practice in this respect is labeling of documents with thesaurus classes for the sake of document classification and management. In the medical domain, for instance, there is a long-standing tradition in terminology maintenance and annotation/classification of documents using standard coding systems such as ICD, MeSH and the UMLS meta-thesaurus. Semantic annotation in a broader sense also addresses document structure (title, section, paragraph, etc.), linguistic structure (dependency, coordination, thematic role, co-reference, etc.), and so forth. In NLP, semantic annotation has been used in connection with machine-learning software trainable on annotated corpora for parsing, word-sense disambiguation, co-reference resolution, summarization, information extraction, and other tasks. A still unexplored but important potential of semantic annotation is that it can provide a common I/O format through which to integrate various component technologies in NLP and AI such as speech recognition, parsing, generation, inference, and so on.

INTELLIGENT CONTENT is semantically structured data that is used for a wide range of content-oriented applications such as classification, retrieval, extraction, translation, presentation, and question-answering, as the organization of such data provides machines with accurate semantic input to those technologies. Semantically annotated resources as described above are typical examples of intelligent content, whereas another major class includes electronic dictionaries and inter-lingual or knowledge-representation data. Some ongoing projects along these lines are GDA (Global Document Annotation), UNL (Universal Networking Language) and SHOE (Simple HTML Ontology Extension), all of which aim at motivating people to semantically organize electronic documents in machine-understandable formats, and at developing and spreading content-oriented application technologies aware of such formats. Along similar lines, MPEG-7 is a framework for semantically annotating audiovisual data for the sake of content-based retrieval and browsing, among others. Incorporation of linguistic annotation into MPEG-7 is in the agenda, because linguistic descriptions already constitute a main part of existing metadata. In short, semantic annotation is a central, basic technology for intelligent content, which in turn is a key notion in systematically coordinating various applications of semantic annotation. In the hope of fueling some of the developments mentioned above and thus promoting the linkage between basic researches and practical applications, the workshop invites researchers and practitioners from such fields as computational linguistics, document processing, terminology, information science, and multimedia content, among others, to discuss various aspects of semantic annotation and intelligent content in an interdisciplinary way.

Paul Buitelaar, Kôiti Hasida

# Organisation and Invited Talks

## Organisers

Paul Buitelaar  
Kôiti Hasida

## Program Committee

Amit Bagga, GE Corporate R&D, USA  
Paul Buitelaar, DFKI-LT, Germany (Co-Chair)  
Gregor Erbach, FTW, Austria  
Christiane Fellbaum, Princeton University, USA  
Wolfgang Giere, ZINFO, University of Frankfurt, Germany  
Nicola Guarino, Ladseb-CNR Padova, Italy  
Kôiti Hasida, ETL, Japan (Co-Chair)  
Boris Katz, AI Laboratory, MIT, USA  
Adam Kilgarriff, University of Brighton, UK  
Elizabeth Liddy, Syracuse University, USA  
Katashi Nagao, IBM TRL, Japan  
Hiroshi Nakagawa, University of Tokyo, Japan  
Hwee Tou Ng, DSO, Singapore  
Martha Palmer, University of Pennsylvania, USA  
Virach Sornlertlamvanich, NECTEC, Thailand  
Steffen Staab, University of Karlsruhe, Germany  
Henry Thompson, Edinburgh University, UK  
Hiroshi Uchida, United Nations University, Japan  
Rémi Zajac, CRL, New Mexico State University, USA

## Invited Talks

*In-depth Utilization of Natural Language Processing for Rich Semantic Annotation*  
Elizabeth Liddy, Syracuse University, USA  
*MindNet as a Framework for Semantically Structuring Text*  
Bill Dolan, Microsoft, USA  
*UNL: Interlingua as Intelligent Content*  
Hiroshi Uchida, United Nations University, Japan  
*GDA: Semantically Annotated Documents as Intelligent Content*  
Koiti Hasida, ETL, Japan

# Table of Contents

Topics and Motivation .....	iii
Organisation and Invited Talks .....	iv
Table of Contents .....	v

## Workshop Papers

### **SECTION 1 : Semantic Annotation of Word Class and Dependency Structure .....** 1

<i>Semantic Annotation of a Japanese Speech Corpus</i> John Fry, Francis Bond .....	3
<i>Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web</i> Eneko Agirre, David Martinez .....	11
<i>Improving Natural Language Processing by Linguistic Document Annotation</i> Hideo Watanabe, Katashi Nagao, Michael McCord, Arendse Bernth .....	20
<i>Building an Annotated Corpus in the Molecular-Biology Domain</i> Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata, Jun-ichi Tsujii .....	28

### **SECTION 2 : Semantic Annotation of Discourse Structure .....** 35

<i>Semantic Annotation for Generation: Issues in Annotating a Corpus to Develop and Evaluate Discourse Entity Realization Algorithms</i> Massimo Poesio .....	37
<i>An Environment for Extracting Resolution Rules of Zero Pronouns from Corpora</i> Hiromi Nakaiwa .....	44
<i>Discourse Structure Analysis for News Video</i> Yasuhiko Watanabe, Yoshihiro Okada, Sadao Kurohashi, Eiichi Iwanari .....	53

### **SECTION 3 : Semantic Annotation of Document Segments .....** 61

<i>Alignment of Sound Track with Text in a TV Drama</i> Seigo Tanimura, Hiroshi Nakagawa .....	63
<i>Semantic Transcoding: Making the WWW More Understandable and Usable with External Annotations</i> Katashi Nagao, Shingo Hosoya, Yoshinari Shirai, Kevin Squire .....	70
<i>From Manual to Semi-Automatic Semantic Annotation: About Ontology-Based Text Annotation Tools</i> Michael Erdmann, Alexander Maedche, Hans-Peter Schnurr, Steffen Staab .....	79