Natural Language Processing (NLP) deals with computational techniques to fully automate any kind of human-like processing of natural language content. The field of NLP has witnessed a phenomenal growth of interest in recent years, both in the academic research space and in the industry scene; the NLP research and development community is thriving today. Important applications of NLP include Machine Translation, Question-Answering, Information Retrieval, Information Extraction, Summarization, Dialogue Systems, Speech Processing, Word Sense Disambiguation, Sentiment and Emotion Analysis, Text Mining, etc. A comprehensive account of all NLP methodologies and applications is beyond the scope of any textbook. In this edited volume, some of the important applications of NLP have been discussed.

It has been observed during several years of teaching and research and development in NLP that some areas of NLP are slowly maturing themselves. These research topics have attracted huge interests, resulted in new workshops and evaluation tracks, and contributed significantly in terms of research publications in recent NLP conferences and journals. However, there are hardly any books collating related readings on such research topics, other than journal publications, workshop / conference proceedings, and evaluation track notes. This book aims to provide relevant theoretical frameworks and the latest empirical research findings in the emerging areas of NLP. Future research directions along with an extensive reference sections on each research topic have also been included.

This edited volume is intended for established NLP researchers who want to improve their understanding of the state-of-the-art in NLP and broaden their research spheres, as well as information-hungry budding researchers looking for related information. The main beneficiary will be the postgraduate and doctoral students’ community who have just stepped into the world of NLP research and are looking for new research problems. A basic background in computational linguistics is expected.

The volume has been divided into six sections. Section 1 contains chapters on text analytics. Section 2 has been devoted to machine translation. Section 3 deals with advanced question answering, while Section 4 presents works on multilingual information access. Digital content management is discussed in Section 5. Finally, Section 6 is on speech processing.

Alexander Gelbukh and Olga Kolesnikova present a survey of contemporary NLP research on Multiword Expressions (MWEs) in Chapter 1. MWEs pose a huge problem to precise language processing due to their idiosyncratic nature and diversity of their semantic, lexical, and syntactical properties. The chapter begins by considering MWE definitions, describes some MWE classes, indicates problems MWEs generate in language applications and their possible solutions, presents methods of MWE encoding in dictionaries and their automatic detection in corpora. The chapter goes into more detail on a particular MWE class called Verb-Noun Constructions (VNCs). Due to their frequency in corpus and unique characteristics, VNCs present a research problem in their own right. Having outlined several ap-
proaches to VNC representation in lexicons, the chapter explains the formalism of Lexical Function as a possible VNC representation. Such representation may serve as a tool for VNCs automatic detection in a corpus. The latter is illustrated on Spanish material applying some supervised learning methods commonly used for NLP tasks.

In Chapter 2, Pushpak Bhattacharyya and Mitesh Khapra present the basic concepts of Word Sense Disambiguation (WSD) and discuss different approaches to solving this problem. The chapter discusses both general purpose WSD and domain specific WSD. The first part of the chapter focuses on existing approaches to WSD, including knowledge-based, supervised, semi-supervised, unsupervised, hybrid, and bilingual approaches. The latter part of the chapter presents a greedy neural network inspired algorithm for domain specific WSD and compares its performance with other state of the art algorithms for WSD.

In Chapter 3, Miriam Lúcia Domingues and Eloi Luiz Favero discuss different approaches to tackling the problem of domain adaptation for part-of-speech tagging and present a case study that has achieved significant accuracy rates on tagging journalistic and scientific texts.

Lucia Specia provides a comprehensive account of Statistical Machine Translation (SMT) in Chapter 4. The chapter introduces the rationale behind SMT, describes the state-of-the-art approaches in SMT, and presents a number of emerging approaches (tree-based SMT, discriminative SMT). The chapter starts with the word-based approach to SMT that forms the basis for more advanced approaches. The general phrase-based approach is discussed along with a number of components that are commonly used within this approach and the standard algorithms used to perform translation (decoding) as well as to estimate the weights of the model components (tuning). Then the chapter presents more recent models that use linguistic information at different stages of phrase-based approaches, including pre- and post-processing stages. The chapter also describes more advanced hierarchical and syntax-based and discriminative models. It also gives an overview of popular metrics for automatic evaluation of SMT systems. The chapter concludes by providing a number of future research directions in this field.

Michael Carl outlines a computational framework for a cognitive model of human translation in Chapter 5. The author investigates the structure of the translators’ keystrokes and gaze data, discusses possibilities for their classification and visualization, and explains how a translation model can be grounded and trained on the empirical user activity data. This chapter starts by outlining the current translation process theories. Then it describes the translators’ user activity data. The chapter then introduces translation progression graphs, a visualization method, and discusses the coarse structure of human translation processes. The next section looks into more details of the user activity data, their segmentation, and classification. The chapter then provides an overview of cognitive architectures. Two models of human translation processes, one based on ACT-R and the other a statistical approach, are then discussed.

In Chapter 6, Patrick Saint-Dizier presents a paradigm for advanced question answering, which mainly includes how-to, why, evaluative, comparative, and opinion questions. These types of questions require quite a lot of discourse semantics analysis and domain knowledge. The first part of the chapter discusses the different parameters at stake in answer production, involving several aspects of cooperation. The second part discusses the text semantics aspects relevant for answering questions. The last part of this chapter introduces <TextCoop>, a platform for discourse semantics analysis for answering complex questions, in particular how-to and opinion questions.

Natalia Konstantinova and Constantin Orasan present the state-of-the-art in the field of Interactive Question Answering (IQA) in Chapter 7. The chapter also briefly presents dialogue systems and question answering, since IQA inherits a lot of features from them. The chapter starts with a description of the characteristics of human dialogues and the basic concepts behind dialogue systems, followed by a brief
introduction to question answering, focusing on the concepts that are used in IQA. Then it presents the most important approaches used by the IQA systems and the challenges in evaluating such systems. The chapter finishes with a description of some large projects, which developed and integrated IQA features.

A large variety of research challenges related to advanced question answering have been presented in Chapter 6. The author focuses on mainly two emerging areas, i.e., developing informative and cooperative answers and using accurate discourse analysis in order to be able to identify well-formed text portions, which are the answer to those questions. These two areas will remain quite challenging for a number of years, in spite of the rapid evolution of the language processing technology. The main focus in Chapter 7 is on Interactive Question Answering (IQA) that has emerged as a research field at the intersection of question answering and dialogue systems, and which allows users to find the answers to questions in an interactive way. The dialogue system plays an important role in IQA, and it initiates dialogues with the user in order to clarify missing or ambiguous information, or suggest further topics for discussion.

Vasudeva Varma and Aditya Mogadala present an account of the issues and challenges in building multilingual information access systems in Chapter 8. The chapter starts with a discussion highlighting the importance of cross-lingual and multilingual information retrieval and information access. The chapter then discusses the distinctions between Cross-Lingual Information Retrieval (CLIR), Multilingual Information Retrieval (MLIR), Cross-Lingual Information Access (CLIA), and Multilingual Information Access (MLIA). The authors subsequently outline issues and challenges in these areas and various approaches, including machine learning-based and knowledge-based approaches, to multilingual information access. It also describes various components of an MLIA system ranging from query processing to output generation. Then evaluation aspects of the MLIA and CLIA systems are discussed at the end of the chapter.

In Chapter 9, Víctor Peinado, Álvaro Rodrigo, and Fernando López-Ostenero survey Multilingual Information Access (MLIA), a multidisciplinary area which aims to solve accessing, querying, and retrieving information from heterogeneous information sources expressed in different languages. The chapter starts by presenting the idea of an Information Retrieval system supporting MLIA, breaking up the three different stages of the Cross-Language Information Retrieval system, namely: (1) processing and indexing the document collection; (2) translation and techniques to overcome the language gap; and (3) matching queries and documents. It also provides details about the difficulties and problems when dealing with multiple languages. The subsequent section focuses on Multilingual Question Answering, a more sophisticated form of IR systems, along with the most successful cross-lingual approaches reported in the field. Then the authors present interactive information retrieval from the user’s perspective with the difficulties associated to conduct and evaluate user-centered experiments: the most relevant results on interactive TREC and CLEF, along with an introduction on user-generated search logs analysis.

While Chapter 8 discusses the issues and challenges in MLIA, Chapter 9 focuses on system architectures that perform Information Retrieval involving MLIA. Chapter 9 also presents Multilingual Question Answering and Interactive Information Retrieval to satisfy users’ information needs.

In Chapter 10, Rafael E. Banchs and Carlos G. Rodríguez Penagos present a detailed introductory account of the most relevant applications of text mining and natural language processing technologies evolving and emerging around the Web 2.0 phenomenon along with the main challenges and new research opportunities that are directly and indirectly derived from them. The chapter starts with providing definitions and general discussions on social media and user-generated content analysis. Some fundamental issues regarding Web 2.0, social media, and natural language processing technologies are covered. Then it presents the main issues and technical challenges related to processing user-generated content in the
context of the Web 2.0, namely encoding, chatspeak, emoticons, ungrammaticality, normalization, co-
referencing, spamming, multilingualism, communication structure, and user roles. It also discusses the
important point of deciding how much natural language processing is warranted for each task, and when
the emerging regularities in vast amounts of data allow for the use of language-independent statistical
methods. Next, the authors present applications related to user-generated content analysis. These ap-
lications include automatic categorization, document summarization, question answering, dialogue
systems, opinion mining and sentiment analysis, outlier identification and misbehavior detection, and
social estimation and forecasting. Then, the chapter presents a section on future trends and research op-
portunities related to processing and analysis of user-generated content in the context of Web 2.0. The
chapter concludes with a discussion on ethical and legal issues concerning privacy and proprietary rights
that arise when processing and using user-generated content.

Lyne Da Sylva offers a detailed overview of digital library management in the perspective of NLP in
Chapter 11. The chapter describes the issues relating to the task of managing a digital library, explores
various NLP applications, which can be applied to the task, and identifies new research problems related
to these issues.

Preety Singh, V. Laxmi, and M. S. Gaur conclude the volume with a presentation on audio-visual
speech recognition, an emerging research topic, in Chapter 12. The chapter describes in detail the various
steps involved in the processing of audio and visual cues for better recognition. It first discusses speech
processing using audio signal only. Then it concentrates on the various steps involved in visual signal
processing, including face detection, lip segmentation, and extraction of lip features. Subsequently, it
discusses the recognition methods that are usually employed for classification of these two types of
signals, and the fusion of the audio and visual streams. The chapter also provides a brief discussion of
some of the available audio-visual databases. The authors then present their proposed methodology for
a visual speech recognition system. They conclude the chapter by mentioning the factors affecting the
performance of lip-reading.

Each chapter included in this edited volume reports a detailed literature survey and provides informa-
tion about existing techniques, including the state-of-the-art, tools, evaluation tracks, and shared tasks
organized on the specific NLP topic discussed. It is expected that early career researchers will find the
chapters informative and useful and asa gateway to the world of NLP.

Sivaji Bandyopadhyay
Jadavpur University, India

Sudip Kumar Naskar
Dublin City University, Ireland

Asif Ekbal
University of Heidelberg, Germany