Preface

Advanced natural languages are one of several predominant conveniences that support our successful being and significantly distinguish us from animals and plants, not speaking of machines. We can communicate, exchange knowledge, acquire or transmit new information, cooperate, manage, govern, and without such a tool known as language, all those matters would be considerably more difficult, if possible at all. Informally said, our languages – and there are several thousands current ones – have a certain syntax given by sets of rules – grammar –, unfortunately including many exceptions, as well as the contents called semantics. Respecting the grammar rules, the contents can be expressed in various forms like speech or written text. Nowadays, tremendous amount of textual documents, created in natural languages, exists and ceaselessly grows because modern electronic tools like various types of databases enable effortlessly storing, retrieving, and transmitting textual data. Provided that the data contains meaningful contents that may be useful, its processing from the semantic point of view is important but how to cope with such big data volumes? Machines – computers – can (and must) help us even if their “thinking” principles are far from ours.

The idea to process natural languages by machines had been born a very long time before the first really working computer – as we commonly use it today – was constructed and put into operation. Probably one of the first scientists, who considered the idea, was a famous German philosopher and mathematician Gottfried Wilhelm Leibniz in the seventeenth century (Hutchins, 1986). He was perhaps inspired by his invention of a mechanical pinwheel calculator along with his deep interest in linguistics and a possibility to create a universal language, including machine translation. From that time, Leibniz’s thoughts have not lost their significance, which has come alive in 1950 by Alan Turing (and later by Noam Chomsky) plus many others based on the introduction of modern powerful computers.

In addition to the traditional analytic tools like statistics, contemporary natural language processing and computational linguistics research (Clark, et al., 2012) focuses strongly on supervised, unsupervised, and semi-supervised learning algorithms in the area of machine learning (Mitchell, 1997), which is part of artificial intelligence (Rusell & Norvig, 2009). Such algorithms are able to learn from data that was (supervised learning) or was not (unsupervised learning) hand-annotated with the desired answers, or using a combination of annotated and non-annotated data (semi-supervised learning). The need to intelligently process data has led to the rise of new informatics areas like data mining (Gaber, 2010) and, from the natural language point of view, text mining (Weiss et al., 2005). The text mining discipline is the intention of this book.
THE BOOK INTENTION

This book, Modern Computational Models of Semantic Discovery in Natural Language, addresses some problems related to automated semantic processing of electronic textual data written in natural languages, including instances of possible solutions demonstrated by carefully selected typical and applicable examples. Providing reliable results of semantic analysis that discovers meaning of textual documents – or their relevant parts – is what we expect from machines as a consequence of increasingly growing very large text-data volumes. Today, the problem solution utilizes combinations of modern computer algorithms including artificial intelligence, machine learning, statistics, and text mining together with linguistic approaches. Natural language processing is now one of the most topical areas both from the research and application point of view because there are – for example, via the Internet – various textual documents available for a common computer-user in many very different professional areas. Using real demonstration examples and representative problem solutions, a reader can learn some of the current possibilities of one of modern computer applications. Natural language processing using common computers (PC’s, laptops) is a very present task with many possible applications because there exist data resources like the Internet that can provide – for example, as a result of browsing – almost unlimited volumes of data as answers. Many areas enable people to freely write their opinions, meaning, sentiment, or reports simply as plain text using any natural language. After collecting a lot of textual documents, the data may contain valuable information that is waiting for its discovery. Because computers and humans use different methods of reasoning, and due to the extremely large volumes of data, new methods supporting semantic analysis have been developed. This publication aims at presenting the contemporary and applicable research results because the progress in this area is very fast, and hopefully may further inspire the next required progress. Natural language processing and text mining are today standard courses at many advanced universities, thus this book can hopefully also serve as a point of departure.

The target audience of the book includes students, teachers, and researchers from schools and universities as well as professionals from companies or organizations that must deal with data having the form of text written in natural languages. The book aims at problems when due to the very large textual-data volumes humans cannot process them manually from the semantic point of view. In practice, companies and interested persons can use this book as an inspiration or starting point for developing their own particular semantic-analysis tools, for example, applied to business intelligence, filtering and categorizing e-mail messages, analyzing meaning of service/goods customers, revealing significant cores or topics in blogs and discussion groups, and so like, or for considering which software tools could be applied according to their needs in the processing of the big accumulated text-data.

Knowing that English is today the most widely used language, the book editors wished also to demonstrate the computational modeling of semantic discovery in other languages because – as a reader certainly knows well – there are many languages coming from different origins that underwent various progress and development.

ORGANIZATION OF THE BOOK

The book is organized into eleven chapters. A brief description of each of the chapters follows:

In the first chapter, Sentiment Classification: Facebook Statuses Mining in the “Arabic Spring” Era, its author, Jalel Akaichi from Tunis, deals with reverberations sent through social networks by the
notable political event known as Arabic Spring. The echoes left their tracks also on popular Facebook. The huge number of informal messages was posted every instant of the day. Feelings seem to be frequently important in these texts for expressing friendship, showing social support, or as part of online arguments. However, existing sentiment analysis studies tend to be commercially-oriented, designed to identify opinions about products rather than user behaviors and state of minds. In his analysis, the author focuses on the application of text mining and sentiment analysis techniques to analyzing Tunisian users’ statuses updates on Facebook. The chapter describes how to extract useful information about the Facebook users’ sentiment and behavior, especially during the Arabic Spring era. To achieve this task, the study describes a method of performing the sentiment analysis by employing the Support Vector Machine and Naïve Bayes data/text mining tools, as well as applying a combination of more than two features. The output of this work consists, on one hand, on the construction of a sentiment lexicon based on the Emoticons and Acronyms’ lexicons that was developed based on the extracted statuses updates; on the other hand, it consists of the realization of detailed comparative experiments between two mentioned machine learning algorithms, Support Vector Machine and Naïve Bayes, by creating a training model for sentiment classification.

The second chapter named Model of the Empirical Distribution Law for Syntactic and Link Words in “perfect” Texts by its author Pavel Makagonov applies mathematically directed modeling to measuring the perfection of the contents and semantic value of an integrated text using connections with the indicators of perfection in the distribution of content words. The criterion for the content of semantically significant words is the coordination of their “frequency-rank” distribution with the Zipf law for short texts and Zipf-Mandelbrot law for long texts. The verified hypothesis is that a perfect system should have not only perfect distribution of its elements – objects, but also perfect connections between them. A model is suggested in which the degree of the text perfection from the point of view of the quality of connections between significative words is determined by the quality of distribution of syntactic and link words in the “rank-frequency” representation. The search for a mathematical representation of this model was conducted proceeding from the assumption that such a model should be a common one not only for the integrated textual document, but also for any system which is considered harmonious, or “perfect”, that is, a system for the objects of which the Zipf law of distribution holds. As a simplified criterion the ratio of the significant and syntactic words used in the analyzed text and the degree of the closeness of this ratio to the “golden section” is considered.

The third chapter, Extracting Definitional Contexts in Spanish through the Identification of Hyponymy-Hyperonymy Relations, from Olga Acosta, Gerardo Sierra, and César Aguilar, makes a reader familiar with the automatic extraction of hyponymy-hypernymy relations in text corpus, which is an important task in Natural Language Processing. This chapter proposes a method for automatically extracting a set of hyponym-hyperonym pairs from a medical corpus in Spanish, expressed in analytical definitions. This kind of definition is composed by a term (the hyponym), a genus term (the hyperonym), and one or more differentiae, that is, a set of particular features proper to the defined term, for example: conjunctivitis is an infection of the conjunctiva of the eye. Definitions are obtained from definitional contexts, and then sequences of term and genus term. Then, the most frequent hyperonyms are used in order to filter relevant definitions. Additionally, using a bootstrapping technique, new hyponym candidates are extracted from the corpus, based on the previous set of hyponyms/hyperonyms detected.

The fourth chapter, Revealing Groups of Semantically Close Textual Documents by Clustering: Problems and Possibilities, written by František Daňena and Jan Žižka, introduces clustering as a family of algorithms that can be successfully used to organize text documents into groups without prior knowledge
of them. The chapter also demonstrates using unsupervised clustering to group large amount of unlabeled textual data (customer reviews written informally in five natural languages) so it can be used later for further analysis. The attention is paid to the process of selecting clustering algorithms, their parameters, methods of data preprocessing, as well as to the methods of evaluating the results by a human expert with an assistance of computers. The feasibility is demonstrated by a number of experiments with external evaluation using known labels and expert validation with the assistance of a computer. It was found that it was possible to apply the same procedures, including clustering, cluster validation, and detection of topics and significant words to different natural languages with satisfactory results.

Similarly, the following fifth chapter, Semantics-based Document Categorization Employing Semi-supervised Learning, from the same authors, Jan Žižka and František Dařena, presents possibilities of another machine-learning based procedure known as semi-supervised learning. The automated categorization of unstructured textual documents according to their semantic contents plays more and more important role particularly linked with the ever growing volume of such data originating from the Internet. Having a sufficient number of labeled examples, a suitable supervised machine-learning based classifier can be trained. When no labeling is available, an unsupervised learning method (clustering) can be applied, however, the missing label information often leads to worse classification results. This chapter demonstrates an alternative method based on semi-supervised learning when a smallish set of manually labeled examples significantly improves the categorization process in comparison with clustering, and the results approach the supervised learning output. For the illustration, a real-world dataset coming from the Internet is used as the input of the supervised learning (Naïve Bayes, Nearest Neighbor, Support Vector Machines), unsupervised one (special k-Means), and semi-supervised one (self-training, co-training). The results are shown for different number of the starting labeled samples used as initial “seeds” to automatically label the large remaining volume of unlabeled items.

The sixth chapter with the title Natural Language Processing as Feature Extraction Method for Building Better Predictive Models comes from the authors Goran Klepać and Marko Velić. This chapter covers natural language processing techniques and their application to predictive model development. Two case studies are presented. First case describes a project where textual descriptions of various situations in a call center of one telecommunication company were processed in order to predict churn. Second case describes sentiment analysis of business news and describes practical and testing issues in text mining projects. Both case studies depict different approaches and are implemented in different tools. Language of the texts processed in these projects is Croatian which belongs to the Slavic group of languages with more complex morphologies and grammar rules than English. Chapter concludes with several points on the future research possible in this domain.

Chapter seven, Departing the Ontology Layer Cake, has two authors: Abel Browarnik and Oded Maimon. Most Ontology Learning approaches follow a model called the Ontology Learning Layer Cake and share many features such as statistical based information retrieval, machine learning, and data and text mining, making resort to linguistics based techniques for certain tasks. The chapter analyzes ontology learning and its goals, as well as the input one would expect when learning ontologies – peer-reviewed scientific papers in English, papers that undergo quality control. After reviewing the model’s shortcomings, the authors suggest an alternative model based on linguistic knowledge. The suggested model would find the meaning of simple components of text – statements. From these statements it should be easy to derive cases and roles that map the reality as a set of entities and relationships or RDF triples, much in the same way as Entity-relationship diagrams do. An analysis of the time complexity of the suggested ontology learning framework shows that this complexity is constant (O(1)) for a sentence,
and $O(n)$ for an ontology based on n sentences. The conclusion is that the Ontology Learning Layer Cake is not adequate for Ontology Learning from text.

In the eighth chapter, *Semantics of Techno-Social Spaces*, the authors Sergey Maruev, Dmitry Stefanovskiy, and Alexander Trousov present a solution related to the so-called techno-social space. Nowadays, most of the digital content is generated within techno-social systems like Facebook or Twitter, where people are connected to other people and to artefacts such as documents and concepts. These networks provide rich context for understanding the role of particular nodes. It is widely agreed that one of the most important principles in the philosophy of language is Frege’s context principle, which states that words have meaning only in the context of a sentence. This chapter puts forward the hypothesis that semantics of the content of techno-social systems should be also analyzed in the context of the whole system. The hypothesis is substantiated by the introduction of a method for formal modeling and mining of techno-social systems and is corroborated by a discussion on the nature of meaning in philosophy. In addition, the authors provide an overview of recent trends in knowledge production and management within the context of our hypothesis.

The ninth chapter, named *Translational Mismatches Involving Clitics (Illustrated from Serbian ~ Catalan Language Pair)* and written by Jasmina Milićević and Àngels Catena, shows that translation of sentences featuring clitics often poses a problem to machine translation systems. The goal of this chapter is to illustrate, on the material from a Serbian ~ Catalan parallel corpus, a rule-based approach to solving translational structural mismatches between linguistic representations that underlie source- and target language sentences containing clitics. Unlike most studies in this field, which make use of phrase structure formalisms, this one was conducted within a dependency framework, that of the Meaning-Text linguistic theory. The authors start by providing a brief description of Catalan and Serbian clitic systems, then introduce the basics of their framework to finally illustrate Serbian ~ Catalan translational mismatches involving the operations of clitic doubling, clitic climbing, and clitic possessor raising.

Chapter ten, *Machine Translation Within Commercial Companies*, deals with what its name suggests: machine translation from the very practical point of view. The chapter author, Tomáš Hudík, gives a short introduction to machine translation and its use within commercial companies with a special focus on the localization industry. Although machine translation is not a new field, many scientists and researchers are still interested in this area and are frequently coming up with challenges, discoveries, and novel approaches. Commercial companies need to keep track with them, and their research & development departments are making good progress with the integration of machine translation within their complicated workflows as well as minor improvements in the core machine-translation in order to gain a competitive advantage. The chapter describes differences in research within university and commercial environments. Furthermore, there are given the main obstacles in the deployment of new technologies and typical ways in which a new technology can be deployed in a corporate environment.

The concluding chapter number eleven, *A corpus-stylistic approach of the Treatises of Great Athanasius about Idolatry*, brings one more excellent demonstration of the natural language processing possibility, this time directed at a historical textual document from the fourth century. The chapter author, Georgios Alexandropoulos, describes the style of the treatises (Contra Gentes and De incarnatione verbi) using corpus analysis tools. This research focuses on certain elements as the most frequently used words, the use of adjectives, pronouns, verbs, lexical bundles, and the most frequently intertextualistic sources. In this interdisciplinary study, classical texts are approached through linguistic tools and the main purpose is to describe the style of a historical person, Great Athanasius, or Saint Athanasius of Alexandria, Egypt – the twentieth bishop in the Coptic Orthodox Church, in these treatises, after having extracted
Preface

all these quantitative data utilizing computational tools. The extraction of the certain data provides us with secure results and leads us to understand that these speeches have certain ideological orientation which is reflected on their linguistic choices. In this way, it is confirmed that the language and style are strictly connected with the ideological intentionality of Great Athanasius, who tries to persuade his audience about the rotten background of idolatry promoting and emphasizing positive things about his religious ideology.

SUMMARY

The presented book demonstrates selected recent attempts to use modern computers as well as advanced scientific tools from the areas like informatics, mathematics, linguistics, machine learning, and artificial intelligence, for mining information and knowledge hidden in real-world textual data-sets that visibly overflow our lives. The intention was to show various applications of modern computational tools to natural language processing without being limited to only one language. Authors from ten countries (Canada, Chile, Croatia, Czech Republic, Greece, Israel, Mexico, Russia, Spain, Tunisia) in eleven chapters presented topical and current points of view to several recent problems and their possible solutions, often as case studies. This book hopes to contribute to the area of the computer-based natural language processing by demonstrating how theory can be used in practice. Also, the editors hope that the book contents can attract new researchers and users of recent, constantly developing area connecting humans and machines.

Jan Žižka
Mendel University in Brno, Czech Republic

František Dařena
Mendel University in Brno, Czech Republic

REFERENCES


