Book Review

Introduction to Linguistic Annotation and Text Analytics

Graham Wilcock
(University of Helsinki)

Princeton, NJ: Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 2, No. 1), 2009, x+149 pp; paperbound, ISBN 978-1-59829-738-6, $40.00; ebook, ISBN 978-1-59829-739-3, $30.00 or by subscription

Reviewed by
Udo Hahn
Friedrich-Schiller-Universität Jena

Looking for a book that lucidly sorts out XML, XSLT and XMI, GATE and UIMA, WordFreak, OpenNLP, and the Stanford NLP Toolkit? Looking for meticulous guidance via batteries of stylesheets and shell scripts, but also keen on exploiting special-purpose rule languages such as JAPE, or pre-configured text analytics engines such as ANNIE? Looking for a coherent set of exercises (covering different technical angles and varying in task complexity) and a series of illustrative screenshots that break down the understanding of annotation workflows into easy-to-digest atomic thematic chunks? Preferring the hands-on "how-to" over dry and winding theory debates and formally based methodological discussions? Interested to get in touch (again) with Shakespeare's Sonnet 130 (the text example running throughout the entire book)? Well, Graham Wilcock's *Introduction to Linguistic Annotation and Text Analytics* might be a perfect match to enjoy all this.

This volume, the third published lecture in Morgan & Claypool's *Synthesis Lectures on Human Language Technologies* series, consists of six chapters. The first one lays the XML-focused meta language foundations and provides additional insights into XML parsing and validation tools, as well as format-switching XML transformation routines using XSL Transformations (XSLT). Chapter 4 continues this technical thread as it elaborates on frameworks for interchanging annotations between different formats using XSLT, as well as the UML-based XML Metadata Interchange (XMI) format, an emerging standard to support the interchange of annotations produced by different tools. Introducing the WordFreak annotation tool, the second chapter sheds light on relevant linguistic annotation layers ranging from formal sentence splitting and tokenization via part-of-speech tagging, syntactic constituency parsing, and semantic predicate–argument analysis up to discourse phenomena such as co-reference resolution. Although this chapter focuses primarily on *manual* annotation (there is also a subsection at the end where WordFreak is linked with OpenNLP), the following one features *automatic* statistical annotation procedures. Easy-to-plug-in OpenNLP is contrasted here with stand-alone Stanford NLP *tools* at all major levels of the linguistic food chain, namely, sentence and token splitting, chunking and parsing, named entity recognition, and co-reference resolution. Increasing the level of abstraction at the systems level, the author then advances to comparing GATE and UIMA, two alternative *architectures* for text analytics. His emphasis is on the proper configuration and task-specific customization (e.g., by introducing the JAPE rule language in GATE and the regex annotator in UIMA, both serving to improve named entity recognition). Again, practical exercises

are discussed in detail running through all levels of linguistic processing (integrating gazetteers/dictionaries, POS tagging, NP chunking and full parsing, named entity recognition, co-reference resolution, etc.). The final chapter concludes with a survey of commercially available tools doing text analytics (e.g., alias-i's LingPipe or IBM's LanguageWare). Furthermore, an advanced treatment of named entity recognition (for job titles) and co-reference resolution is provided using the open-source frameworks of rule-based GATE and UIMA solutions (both incorporating customized dictionaries) and statistically based OpenNLP.

The book assumes little prior knowledge, although regular expression, JAVA, and XML basics will certainly be helpful. It is targeted at undergraduate level (not necessarily computer science) students who wish to gather experience in reusing, modifying, and customizing existing linguistic annotation tools and text analytics architectures. Throughout the book, the author directs the reader to open-source, freely available, platform-independent, and easily downloadable software resources and repositories. So students find themselves embedded in a stimulating playground where tooling is the message.

The book is comprehensively written, well-structured, and very easy to follow, in particular when students have the exercises run simultaneously on their personal machines. The author has a clearly designed didactical master plan in mind which is realized by a large number of exercises with increasing, but never particularly high, complexity (see, e.g., the fourth chapter that deals with a nice set of transformation problems involving WordFreak–OpenNLP [XML–plain text], GATE–WordFreak [XML–XML], and WordFreak–UIMA [XML–XMI] tool [language] pairs). These exercises are reasonably chosen and solutions are technically well prepared and exhaustively explained with an admirable degree of clarity.

However, the more advanced and experienced reader might find the continuous pampering by way of overly fine-grained thematic exposition and visualization a bit excessive, perhaps even wearisome. There is, for example, also no division into easy, medium, and hard problems to offer challenging tasks at different levels of students' understanding. If the book is used in the context of more advanced teaching, it should be complemented by much more technical standard reference textbooks providing complementary theoretical background information on empirical NLP, (supervised) machine learning methods for NLP, computational corpus linguistics, and so forth. Yet, for getting used to mapping routines, workflows, and software underlying linguistic meta data annotation, this tutorial fills certainly a gap for students who come across these topics for the first time and enjoy the all-embracing tutorial approach of the author.

Any real complaints? Just a minor remark: all (URL) references are almost unreadable (Web links were not removed from the printed version of the book).

There is also a dedicated Web site which contains copies of the material from this book.[1]

This book review was edited by Pierre Isabelle.

Udo Hahn is Professor of Computational Linguistics and Language Technology at Friedrich-Schiller-Universität Jena (Germany) and Head of the Jena University Language & Information Engineering (JULIE) Lab. He works on text analytics (semantic search technologies, information extraction, text mining, and text summarization), with focus on biomedical language processing. Hahn's address is Computerlinguistik, Friedrich-Schiller-Universität Jena, D-07743 Jena, Germany; e-mail: udo.hahn@uni-jena.de; URL: http://www.julielab.de/.

1 http://sites.morganclaypool.com/wilcock.