Web Corpus Construction

Synthesis Lectures on Human Language Technologies

Editor
Graeme Hirst, ***University of Toronto***

Synthesis Lectures on Human Language Technologies is edited by Graeme Hirst of the University of Toronto. The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Web Corpus Construction
Roland Schäfer and Felix Bildhauer
2013

Recognizing Textual Entailment: Models and Applications
Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing
Anders Søgaard
2013

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
Emily M. Bender
2013

Semantic Relations Between Nominals
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative
Inderjeet Mani
2012

Natural Language Processing for Historical Texts
Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining
Bing Liu
2012

Discourse Processing
Manfred Stede
2011

Bitext Alignment
Jörg Tiedemann
2011

Linguistic Structure Prediction
Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing
Hang Li
2011

Computational Modeling of Human Language Acquisition
Afra Alishahi
2010

Introduction to Arabic Natural Language Processing
Nizar Y. Habash
2010

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

Web Corpus Construction
Roland Schäfer and Felix Bildhauer

ISBN: 978-3-031-01024-8 paperback
ISBN: 978-3-031-02152-7 ebook

DOI 10.1007/S00508ED1V01Y201305HLT022

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #22
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Synthesis Lectures on Human Language Technologies
Print 1947-4040 Electronic 1947-4059

Web Corpus Construction

Roland Schäfer and Felix Bildhauer
Freie Universität Berlin

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #22

ABSTRACT

The World Wide Web constitutes the largest existing source of texts written in a great variety of languages. A feasible and sound way of exploiting this data for linguistic research is to compile a static corpus for a given language. There are several advantages of this approach: (i) Working with such corpora obviates the problems encountered when using Internet search engines in quantitative linguistic research (such as non-transparent ranking algorithms). (ii) Creating a corpus from web data is virtually free. (iii) The size of corpora compiled from the WWW may exceed by several orders of magnitudes the size of language resources offered elsewhere. (iv) The data is locally available to the user, and it can be linguistically post-processed and queried with the tools preferred by her/him.

This book addresses the main practical tasks in the creation of web corpora up to giga-token size. Among these tasks are the sampling process (i. e., web crawling) and the usual cleanups including boilerplate removal and removal of duplicated content. Linguistic processing and problems with linguistic processing coming from the different kinds of noise in web corpora are also covered. Finally, the authors show how web corpora can be evaluated and compared to other corpora (such as traditionally compiled corpora).

For additional material please visit the companion website

`http://sites.morganclaypool.com/wcc`

KEYWORDS

corpus creation, web corpora, web crawling, web characterization, boilerplate removal, language identification, duplicate detection, near-duplicate detection, tokenization, POS tagging, noisy data, corpus evaluation, corpus comparison, keyword extraction

Contents

Preface

Our approach to the subject of web corpus construction is guided by our own practical experience in the area. Coming from an empirically oriented linguistics background, we required large amounts of data for empirical research in various languages, including more or less non-standard language. However, we noticed that, depending on the research question and the language of interest, appropriate text resources are not always available and/or freely accessible and in the appropriate form (cf. Section 1 for examples). Therefore, we took the work by the WaCky initiative [Baroni et al., 2009] and the Leipzig Corpora Collection (LCC, Biemann et al., 2007; Goldhahn et al., 2012) as a starting point to build our own large corpora from web data, leading to the development of the `texrex` software suite and the COW ("COrpora from the web") corpora.[1,2]

We dealt with the usual technical problems in web corpus construction, like boilerplate removal and deduplication, noticing that there was no concise and reasonably complete introductory textbook on these technicalities available, although there are overview articles like Fletcher [2011]; Kilgarriff and Grefenstette [2003]; Lüdeling et al. [2007]. Additionally, it became clear to us that even the mere use of web corpora for linguistic research requires extra precautions and more in-depth knowledge about the corpus construction process compared to the use of established and "clean" corpus resources. This knowledge—mostly specific to web corpora—includes important matters like:

- How was the corpus material sampled, which in this case means "crawled"?
- Which parts of the documents are removed in the usual "cleaning" steps, and with which accuracy?
- Which documents are removed completely by which criteria, for example, near-duplicate documents?
- What kinds of noise are present in the data itself (e. g., misspellings), and what was normalized, removed, etc., by the corpus designers?
- Which kinds of noise might be introduced by the post-processing, such as tokenization errors, inaccurate POS tagging, etc.?

The literature on these subjects comes to some extent (or rather to a large extent) from the search engine and data mining sphere, as well as from Computational Linguistics. It is also quite diverse, and no canonical set of papers has been established yet, making it difficult to get a complete picture in a short time. We hope to have compiled an overview of the papers which can be considered recommended readings for anyone who wishes to compile a web corpus using their

[1] `http://sourceforge.net/projects/texrex/`
[2] `http://www.corporafromtheweb.org/`

own tools (own crawlers, boilerplate detectors, deduplication software, etc.) or using available tools.[3] Equally important is our second goal, namely that this tutorial puts any web corpus user in a position to make educated use of the available resources.

Although the book is primarily intended as a tutorial, this double goal and the extremely diverse background which might be required leads to a mix of more practical and more theoretical sections. Especially, Chapter 2 on data collection contains the least amount of practical recommendation, mainly because data collection (primarily: web crawling) has—in our perception—received the least attention (in terms of fundamental research) within the web corpus construction community. Chapters 3 on non-linguistic post-processing and 4 on linguistic post-processing are probably the most practical chapters. Chapter 5 briefly touches upon the question of how we can assess the quality of a web corpus (mainly by comparing it to other corpora). Thus, it is of high theoretical relevance while containing concrete recommendations regarding some methods which can be used.

Roland Schäfer and Felix Bildhauer
July 2013

[3] We make some software recommendations, but strictly from the open source world. We do this not so much out of dogmatism, but rather because there are open source variants of all important tools and libraries available, and nobody has to pay for the relevant software.

Acknowledgments

Much of the material in this book was presented as a foundational course at the European Summer School in Logic, Language and Information (ESSLLI) 2012 in Opole, Poland, by the authors. We would like to thank the ESSLLI organizers for giving us the chance to teach the course. We also thank the participants of the ESSLLI course for their valuable feedback and discussion, especially Ekaterina Chernyak (NRU-HSE, Moscow, Russia). Also, we are grateful for many comments by participants of diverse talks, presentations, and workshops held between 2011 and 2013. Furthermore, we would like to thank Adam Kilgarriff and two anonymous reviewers for detailed and helpful comments on a draft version of this book. Any errors, omissions, and inadequacies which remain are probably due to us not listening to all these people.

We could not have written this tutorial without our prior work on our own corpora and tools. Therefore, we thank Stefan Müller (Freie Universität Berlin) for allowing us to stress the computing infrastructure of the German Grammar work group to its limits. We also thank the `GNU/Linux` support team at the *Zedat* data centre of Freie Universität Berlin for their technical support (Robert Schüttler, Holger Weiß, and many others). Finally, we thank our student research assistant, Sarah Dietzfelbinger, for doing much of the dirty work (like generating training data for classifiers).

The second author's work on this book was funded by the *Deutsche Forschungsgemeinschaft*, SFB 632 "Information Structure," Project A6.

Roland Schäfer would like to thank his parents for substantial support in a critical phase of the writing of this book.

Felix Bildhauer is very much indebted to Chiaoi and Oskar for their patience and support while he was working on this book.

Roland Schäfer and Felix Bildhauer
July 2013