Web Corpus Construction

Synthesis Lectures on Human Language Technologies

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

www.morganclaypool.com

ISBN: 9781608459834 paperback
ISBN: 9781608459841 ebook

DOI 10.2200/S00508ED1V01Y201305HLT022

A Publication in the Morgan & Claypool Publishers 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

MORGAN & CLAYPOOL PUBLISHERS

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 adavantages 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

CHAPTER 1

Web Corpora

Although corpus-based Linguistics has seen a rise in popularity over the past few decades, for many research questions the available resources are sometimes too small, sometimes too unbalanced, or they are balanced according to inappropriate criteria for the task, sometimes too close to the respective standard language (again, for certain types of research questions), and sometimes they are simply too expensive. Sometimes, it is also the case that the interfaces provided to access available corpora are too restricted in search or export options to do serious quantitative research or use the corpus for Computational Linguistics tasks. In addition, many freely available corpora cannot be downloaded as a whole, which is required for many applications in Computational Linguistics. Examples of the above include:

- The German *Deutsches Referenzkorpus* (DeReKo; Kupietz et al., 2010) by the *Institut für Deutsche Sprache* (IDS) is large (currently over 5 billion words), but it contains predominantly newspaper text and is therefore unsuitable for research which requires a corpus containing a variety of registers, genres, etc.
- The corpus by the *Digitales Wörterbuch der Deutschen Sprache* (DWDS; Geyken, 2006) is a carefully balanced corpus of the German language of the 20th century, optimized for lexicographic research. However, it contains only 123 million tokens. On top of the small size, export options are highly limited, and many texts in the corpus are not licensed for export by non-project members.[1]
- Most of this is true for the British National Corpus (BNC; Leech, 1993).
- The corpora distributed by the Linguistic Data Consortium are small and expensive. They often add huge value through rich annotation, but while this is extremely useful for some types of research, in some areas researchers need corpora several orders of magnitude larger. E. g., the Penn Treebank [Marcus et al., 1993, 1999] contains roughly 4.5 million words and costs $3,150 at the time of this writing according to the web page.[2]
- The French Frantext corpus, provided by *Analyse et Traitement Informatique de la Langue Française* (ATILF), is a collection consisting predominantly of fictional texts and philosophical literature.[3,4] As of early 2013, it comprises slightly over four thousand documents

[1] Since any criticism regarding the specific composition of balanced corpora is guided by individual scientific needs and therefore futile, we will not engage in it.
[2] http://ldc.upenn.edu/
[3] http://www.frantext.fr/
[4] http://www.atilf.fr/

(the providers do not publish any token counts on the website), ranging from the 12th to 21st century (and including 850 texts from 1950 or later).

- The Spanish *Corpus de Referencia del Español Actual* (CREA) by *Academia Real Española* contains 155 million tokens and is a balanced corpus of predominantly written language from many different Spanish-speaking countries. Access through a WWW interface is free, but the query language is rather limited, there are bugs in the query engine, the display options are limited, and query results cannot be exported.
- The Spanish *Corpus del Español* [Davies, 2002] is offered with an advanced query interface and contains texts from the 13th to the 20th centuries which sum up to 100 million word tokens. However, contemporary Spanish (20th century) is represented by a mere 20 million tokens.
- The Swedish *Språkbanken* project at *Göteborgs Universitet* offers free access to roughly 1 billion tokens through a web interface.[5] It is thus quite large. However, it cannot be downloaded as a whole.

In Theoretical Linguistics, researchers sometimes try to obviate limitations of available resources through Googleology. Especially when data on low-frequency or non-standard phenomena is needed, search engine queries (mostly using Google's service) are used to look for single occurrences of some grammatical construction, or—even worse—result counts returned for such queries are used for more or less formal statistical inference. This must be considered bad science by any account. Everything that needs to be said about Googleology was already said in Kilgarriff [2006]. Problems with Googleology include, but are not restricted to:

1. Search engines are designed to favor precision over recall [Manning et al., 2009, 432] according to non-linguistic relevance criteria. This means that we can never be sure that we have found all or even some of the linguistically relevant results.
2. The ranking criteria are not known exactly, although usually variants of PageRank-like measures [Brin and Page, 1998] are used (cf. Section 2.4.2). Ranking can even be influenced by economical factors (sponsored ranking). For good research practice, some form or random sampling from the corpus would be required.
3. Search engines adapt the ranking of the results according to information like the language of the browser or the headers sent by the client which inform the server about the user's preferred language, geo-location of the client's IP address, etc. In practice, this means that two identical queries almost never result in the same results being displayed, which makes any claim based on these results non-reproducible—a clear indication of bad science.
4. Search engines expand and reduce search terms without providing feedback about the exact nature of the expansions and reductions. This includes methods like spelling correction and morphological analysis [Cucerzan and Brill, 2004]. Queries are also optimized, e. g., by bracketing sub-expressions of the original query in order to improve the overall precision.

[5] http://spraakbanken.gu.se/

Such methods are often based on analyses of query logs [Guo et al., 2008; Hagen et al., 2011; Risvik et al., 2003]. Especially if we take query result counts to make quantitative statements, we do not know even remotely what kind of (expanded or reduced) queries they actually represent.

5. Despite performing a lot of such covert linguistic processing, search engines offer no linguistic annotation. Controlled wildcarding (including wildcarding over parts-of-speech or lemmata) is not available. This means that we cannot apply the usual search heuristics required to find all relevant examples in a corpus.
6. Search engines return estimated total counts, which fluctuate on a daily basis. A few papers, such as Eu [2008] and recently Rayson et al. [2012] discuss this fluctuation and its impact on linguistic research. In Rayson et al. [2012], a method is suggested to derive stable frequencies from fluctuating search engine counts through long-term observation. It is stated that more advanced time series statistics could provide even better results. The counts would still depend on which documents the search engine considers worth indexing, and the practical feasibility compared to the use of a static web corpus needs to be proven.
7. In any case, counts returned by search engines are usually page counts, not token counts. Whether the number of web pages on which a term or expression occurs is relevant at all is a very open question (but see Keller and Lapata, 2003, who find a high correlation between the counts returned by search engines and frequencies in traditional corpora like the BNC).
8. The "corpus" which is indexed by a search engine (some part of the indexable web deemed relevant by the search engine provider) changes frequently, including the massive creation of new pages and the removal of old ones. To deal with this, the index of a search engine is constantly updated, but different pages are re-indexed at quite different intervals. This means that pseudo-scientific results obtained from search engines cannot be reproduced by other researchers. Even if all other points of criticism put forward here were invalid, this single fact makes reporting search engine results bad science.

To avoid such problems with search engine results and to fill the gaps left by available traditionally compiled corpora, web corpora as popularized by the WaCky initiative [Baroni et al., 2009] or the Leipzig Corpora Collection (LCC, Biemann et al., 2007) are an ideal solution. If available traditional or web corpora do not suffice, then a web corpus can in principle be compiled by any researcher ad hoc and according to specific design decisions. The Leipzig Corpora are available in many languages, but for legal reasons, only sentence-wise shuffled versions are published, which makes them inappropriate for any research at the document-level. The WaCky corpora are considerably larger and contain whole documents, but they are only available for a handful of European languages. In some cases, even larger corpora than the WaCky corpora might be required.

To summarize, the main advantages of web corpora are:

1. They can be constructed at no cost beyond expenses for hard drive space, CPU power, and bandwidth.

2. They can reach almost arbitrary sizes, far in the giga-token region.
3. Entirely new registers and genres are available exclusively in web corpora (blogs, forums, etc.). Some of these genres cover texts which are much closer to spontaneous (although not necessarily spoken) language than texts contained in traditionally compiled corpora.
4. They are available locally for all kinds of processing using standard tools, database systems, etc., such that the user is not bound by the limitations of some query interface.
5. web corpus construction allows for a form of random sampling from the population of web documents (even uniform sampling, where each document has the same chance of being sampled). Furthermore, the population of web documents is very large and extremely diverse in terms of genres, styles, etc. Constructing truly random samples from such a huge and diverse population is not possible by means of traditional corpus construction. Corpora constructed in such a way could therefore be a valuable addition to available resources based on stratified sampling (i. e., balanced corpora).

The advantages have to be balanced against the disadvantages:

1. The copyright situation is unclear in some countries where there is no fair use policy (like Germany). Especially the redistribution of the corpora is problematic under such conditions.
2. Although there are advanced methods to clean web corpora, they still contain significant amounts of noisy data, such as spam web pages, redundant auto-generated content from content management systems, misspellings, etc. Compared to users of traditional corpora, users of web corpora must therefore be more aware of the steps which were taken in the construction of the corpus, such that they are aware of potential distortions of their results. As an example which will be explained in detail in Chapters 4 and 5 the word type count for web corpora is usually much too high to be plausible due to a large amount of noisy material, such that naïvely drawn statistical conclusions might be invalid.

Since web corpora can serve many different purposes, it is probably impossible to satisfy everyone with one single introductory text book. We look at the subject mainly from the point of view of empirical linguistics, which is why we kept some sections shorter than computational linguists probably would have. For example, this is certainly true for Section 2.4.3 about focused crawling (the task of mining the web for documents about certain topics, in certain languages, etc.). In making such decisions, we mainly adopted the following guideline in order to stay compatible with non-computational linguists: If there are available tools which solve a task without requiring that the user have advanced programming skills, we deal with it in detail, otherwise we discuss it briefly and include suggestions for further reading. Creating and running a focused crawler is definitely a task that requires advanced programming skills, and we therefore kept the section about it comparatively short.

Throughout the book, we will be suggesting that certain technical decisions are actually design decisions, because they influence the features of the final corpus. This is especially crucial

because web corpus construction (as understood here) is a process which must usually be fully automated due to the size of the corpora. All automatic processing alters the original data, it comes with a certain error rate, and it might decrease the quality of the data instead of increasing it. This concerns the tasks of removing duplicate documents, removing design elements (menus, copyright strings, etc.) from web pages, and other cleanup and processing procedures. A good example is removal of duplication. If it is performed at the paragraph level instead of the document level, single documents will end up being incomplete in the final corpus. For some researchers, this might simply not be acceptable. Given the diverse applications of web corpora, users have to make the design decisions themselves, based on our description of how the procedures work.

We have divided the book into four main chapters. Each chapter covers a major topic in web corpus construction, and they are ordered in the usual order of practical corpus construction. Chapter 2 describes the theory and technology of data collection (crawling). Chapter 3 provides the non-linguistic cleansing which is usually applied to the collected data. Chapter 4 discusses the problems which arise in the linguistic processing (tokenizing and annotations like POS tagging) of web data. Finally, Chapter 5 introduces some methods of determining the quality of the compiled corpora.