Building and Using Comparable Corpora

Serge Sharoff · Reinhard Rapp
Pierre Zweigenbaum · Pascale Fung
Editors

Building and Using Comparable Corpora

Springer

Editors
Serge Sharoff
Centre for Translation Studies
University of Leeds
Leeds
UK

Reinhard Rapp
University of Mainz
Mainz
Germany

Pierre Zweigenbaum
LIMSI-CNRS
Université de Paris-Sud
Orsay
France

Pascale Fung
Department of Science and Technology,
Electronic and Computer Engineering
The Hong Kong University
Hong Kong
People's Republic of China

ISBN 978-3-642-20127-1 ISBN 978-3-642-20128-8 (eBook)
DOI 10.1007/978-3-642-20128-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013947575

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book came from the experience of a series of annual BUCC workshops. The first workshop of this kind was held in 2008 at LREC in Marrakech organised by Pierre Zweigenbaum, Éric Gaussier and Pascale Fung. Since then, the workshops changed the continents (Singapore in 2009, Malta in 2010, Portland, Oregon, 2011, Istanbul 2012); the organising committee included Reinhard Rapp, Serge Sharoff and Marko Tadic, but its main topic remained the same, focusing on the need to use comparable corpora as training data for linguistic research and NLP applications. The chapters for this volume were collected mostly from the best submissions to the workshops at the end of 2011 or through specific requests to the most prominent authors in this field. After completing the editorial process the collection of chapters is presented to your attention.

The volume starts with a chapter overviewing the state of the art. It discusses the rationale behind the use of comparable corpora, as well the issues involved in their collection, annotation and use. The rest of the volume consists of two parts. Part I is devoted to methods of compiling comparable corpora and measuring the degree of comparability between their documents. Part II is on applications which use comparable corpora in various contexts such as Machine Translation or computer-assisted human translation.

In Part I there are eight chapters.

"Mining Parallel Documents Using Low Bandwidth and High Precision CLIR from the Heterogeneous Web" by Shi and Fung proposes a method for mining parallel documents, which is based on the principles of cross-lingual information retrieval. The quality of resources obtained in this way is evaluated by using SMT.

"Automatic Comparable Web Corpora Collection and Bilingual Terminology Extraction for Specialized Dictionary Making" by Gurrutxaga et al. presents two tools, respectively, for compiling comparable corpora from the Web and for extracting bilingual terminology from them. The authors are specifically interested in under-resourced languages, Basque in their case, when the number of relevant webpages is relatively small. The paper describes the use of the standard tools (Boot- Cat, Kimatu, etc) and application of the context feature vectors for aligning monolingual term lists.

"Statistical Comparability: Methodological Caveats" by Köhler explores the issue of corpus comparability from the viewpoint of statistical testing. It shows how the notions of statistics of frequency distributions, such as homoscedasticity

and skewness, can be applied to analyse comparable corpora, including the issues of their representativeness, homogeneity, as well as comparability.

"Methods for Collection and Evaluation of Comparable Documents" by Paramita et al., similar to "Automatic Comparable Web Corpora Collection and Bilingual Terminology Extraction for Specialized Dictionary Making", also focuses on the collection of comparable corpora from the Web for under-resourced languages. The authors investigate the use of the interwiki links in Wikipedia and retrieval of Twitter tweets by using URLs and topics as queries. They also propose methods to evaluate the retrieved documents using automatic classification of their comparability levels.

"Measuring the Distance Between Comparable Corpora Between Languages" by Sharoff explores methods for comparing corpora of unknown composition using keywords. First, he explores attempts at approximating the content of corpora collected from the Web using various methods, also in comparison to traditional corpora, such as the BNC. The procedure for estimating the corpus composition is based on selecting keywords, followed by clustering. This can apply to corpora within the same language, e.g., the BNC against ukWac as well as to corpora in different languages, e.g., webpages collected using the same procedure for English and Russian.

Li and Gaussier ("Exploiting Comparable Corpora for Lexicon Extraction: Measuring and Improving Corpus Quality") take care of an important property of comparable corpora: their degree of comparability. They propose a measure of comparability which is linked to the possibility of extracting word translations from comparable corpora. They show that this measure correlates with intuition on a range of artificial comparable corpora. They design a bilingual clustering method which increases this measure through a controlled extension of initial comparable corpora, and show that the bilingual lexicons they extract from these corpora are indeed improved by this process.

"Statistical Corpus and Language Comparison on Comparable Corpora" by Eckart and Quasthoff describes the construction of the Leipzig Corpus Collection which currently grows at a rate of 30 GB per month. It also gives an overview on a number of applications of this comparable data, and highlights some of its statistical properties such as n-gram frequencies, word co-occurrences and the distributions of word and sentence lengths. An integral part of the system is a web portal which gives an overview on the corpora and serves as a starting point for evaluating phenomena relating to corpus, genre and language comparison.

"Comparable Multilingual Patents as Large-Scale Parallel Corpora" by Lu and Tsou describes methods used for building a large-scale multilingual corpus of comparable patents for a range of languages, such as Chinese, Japanese, Korean and German. The chapter also discusses a procedure to extract parallel sentences from these patents to build an SMT system.

In Part II we also have eight chapters.

"Extracting Parallel Phrases from Comparable Data" by Hewavitharana and Vogel deals with the problem of discovering parallelness in comparable data at the sub-sentential level, i.e. to extract parallel phrases embedded in comparable

sentences. They explore and quantitatively compare three different approaches: Using the standard Viterbi phrase alignment, using lexical features only without relying on the Viterbi path of word alignments, and using a maximum entropy classifier which is applied on large collections of phrase pair candidates. Their finding is that the second approach leads to the best results in terms of F-measure.

Similar to the first chapter of Part II, the "Exploiting Comparable Corpora", authored by Munteanu and Marcu, also deals with the extraction of parallel subsentential fragments from comparable corpora. However, they use a completely different approach which is based on signal filtering, whereby the signal is derived from word translation probabilities. Also, their evaluation procedure is task based. They showed that by adding the parallel data as extracted from the comparable corpora to the non-domain specific parallel training data of a statistical MT system, the translation quality improved.

Deléger et al. ("Paraphrase Detection in Monolingual Specialized/Lay Comparable Corpora") study same-language comparable corpora, where the dimension of comparability is a contrast in discourse type: texts intended for specialists of a domain (health) versus texts intended for lay people. They identify systematic variations in the expression of information in such comparable corpora. For this purpose, they test both a top-down approach, applying given variation patterns, and a bottom-up approach, discovering such patterns from the observation of data. The most common patterns evidence a preference for verb nominalisations and for relational adjectives in specialized language, as opposed to lay language.

Ji et al. ("Information Network Construction and Alignment from Automatically Acquired Comparable Corpora") describe an approach for acquiring cross-lingual comparable corpora which is based on concept extraction from videos. The corpora thus obtained are then used to identify translations of names using a weakly supervised and language-independent bootstrapping approach. The approach uses as seeds expressions that have the same forms in different languages, and—based on link comparison—iteratively mines more and more name translations.

Morin et al. ("Bilingual Terminology Mining from Language for Special Purposes Comparable Corpora") deal with small comparable corpora (250 kwords) in specialised domains, which reduces the discriminative power of the context vectors used in the standard approach of bilingual lexicon extraction. They propose two directions which they show improve bilingual lexicon extraction in this situation. First, to make them more discriminant, they boost the importance given to specific words which they consider as 'anchor points': transliterated words and neoclassical compounds. Second, they experiment with a small, in-domain parallel corpus from which they extract an additional bilingual lexicon which they use to extend the seed lexicon of the standard method.

Kageura and Abekawa ("The Place of Comparable Corpora in Providing Terminological Reference Information to Online Translators: A Strategic Framework") notice that the recent advances in the term extraction and alignment methods are not taken into translation practice. They are primarily interested

in using comparable corpora to provide terminological resources, especially in the context of online collaborative translation. They advocate the use of comparable corpora for a *posteriori* enquiries after bilingual term candidates have been extracted.

"Old Needs, New Solutions: Comparable Corpora for Language Professionals" by Bernardini and Ferraresi is also concerned about the use of comparable corpora by professional translators. The authors investigate the context for using different types of corpora, including small *ad hoc* corpora (very small, but reliable), large web-derived reference corpora (with abundance of data, but little specialisation) and interactively constructed semi-automatic corpora, which occupy the middle ground and offer a positive trade-off between the effort needed to construct the corpora and their perceived usefulness.

"Exploiting the Incomparability of Comparable Corpora for Contrastive Linguistics and Translation Studies" by Neumann and Hansen-Schirra investigates the use of a comparable corpus of English and German, which includes both monolingually comparable texts and texts with their translations. The chapter provides insights from a feature matrix to reveal differences and commonalities between the original texts in two languages (English and German) as well as between originals and their translations in the same language.

Serge Sharoff
Reinhard Rapp
Pierre Zweigenbaum
Pascale Fung

Acknowledgments

In the process of preparing this volume we received a lot of help from the publishers, especially from Olga Chiarcos and Federica Corradi dell Acqua, as well as from the editors of related series, Ed Hovy and Nancy Ide. We really appreciate their advice. We are also grateful to our authors and reviewers, and for the support obtained from the 7th European Community Framework Programme via the projects TTC, Monotrans, HyghTra, and AutoWordNet.

Serge Sharoff
Reinhard Rapp
Pierre Zweigenbaum
Pascale Fung

Contents