Synthesis Lectures on
Human Language Technologies

SYNTHESIS
COLLECTION OF TECHNOLOGY

Serge Sharoff · Reinhard Rapp ·
Pierre Zweigenbaum

Building and Using Comparable Corpora for Multilingual Natural Language Processing

Springer

Synthesis Lectures on Human Language Technologies

Series Editor

Graeme Hirst, Sandford Fleming Building, Department of Computer Science, University of Toronto, Toronto, ON, Canada

The series publishes 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.

Serge Sharoff · Reinhard Rapp ·
Pierre Zweigenbaum

Building and Using Comparable Corpora for Multilingual Natural Language Processing

Springer

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

Reinhard Rapp
Faculty of Translation Studies, Linguistics
and Cultural Studies
University of Mainz
Germersheim, Germany

Pierre Zweigenbaum
CNRS, Laboratoire Interdisciplinaire des
Sciences du Numérique
Université Paris-Saclay
Orsay, France

ISSN 1947-4040 ISSN 1947-4059 (electronic)
Synthesis Lectures on Human Language Technologies
ISBN 978-3-031-31383-7 ISBN 978-3-031-31384-4 (eBook)
https://doi.org/10.1007/978-3-031-31384-4

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

About the Authors

Serge Sharoff is Professor of Language Technology and Digital Humanities at the Centre for Translation Studies, University of Leeds. Over the course of his academic career, he has published more than 150 peer-reviewed papers and supervised 18 Ph.D. students. His research focuses on Natural Language Processing, including automated methods for collecting very large corpora from the Web, their analysis in terms of domains and genres as well as extraction of lexicons and terminology from corpora. The application domains for this kind of research in the Digital Humanities include text annotation, information retrieval, machine translation and computer-assisted language learning. His research stresses the inherent multilingualism of NLP, which implies that tools and resources can be ported across languages by paying attention to the respective linguistic properties.

Reinhard Rapp is Professor of Applied Translation Studies at Magdeburg-Stendal University of Applied Sciences and is also affiliated with the University of Mainz. His main research interests are in computational linguistics, translation studies and cognitive science. Publications have dealt with unsupervised language learning from text corpora, word sense disambiguation, text mining, thesaurus construction, bilingual dictionary induction from parallel and comparable corpora, and with statistical and neural machine translation. He conducted EU-funded research projects at the University of Geneva, the University of Tarragona, the University of Leeds, at Aix-Marseille University, at the University of Mainz and at the Athena Research Center in Athens. He co-organized about 30 scientific conferences, workshops and shared tasks and is a co-editor of the book series 'Translation and Multilingual Natural Language Processing' (Language Science Press) and 'Linguistik International' (Peter Lang).

Pierre Zweigenbaum Ph.D., FACMI, FIAHSI, is a Senior Researcher at the Interdisciplinary Laboratory for Digital Sciences (LISN, Orsay, France), a laboratory of the French National Center for Scientific Research (CNRS) and Université Paris-Saclay, where he led the ILES Natural Language Processing group for seven years. Before CNRS he was a researcher at Paris Public Hospitals in an Inserm team for twenty years. He also was a part-time professor at the National Institute for Oriental Languages and Civilizations

during ten years. His research focus is Natural Language Processing, with medicine as a main application domain. He has also designed methods to acquire linguistic knowledge automatically from corpora and thesauri, to help extend monolingual and bilingual lexicons and terminologies, using parallel and comparable corpora.

Introduction

1

1.1 Rationale for Working with Comparable Corpora

The ability of the computers to handle larger amounts of texts and the availability of more texts in electronic form led to the rise of data driven research in computational linguistics. In the case of Machine Translation (MT) and other kinds of multilingual Natural Language Processing (NLP), the first source of large data came from collections of translations, initially in the Statistical MT (SMT) approach from IBM [1], which was based on the Proceedings of the Canadian Parliament in English and French as their data source. This research direction was followed by a proliferation of SMT models, which relied on larger and larger collections of *parallel* data, which consist of exact translations between a pair of languages or several languages at the same time.

However, this early research also demonstrated the limits of using fully parallel corpora, with the most pressing initial concern on the amount of such data [4, 5]. This led to another strand of studies concerning the use of *less* parallel sources of texts, usually under the name of comparable corpora [6]. This book will present an overview of the modern approaches to building and using comparable data in multilingual NLP.

We will start by outlining the basic principles of using comparable resources as well as by comparing them to fully parallel resources (this chapter). This will be followed by specific chapters on building comparable corpora (Chap. 3), aligning their sentences to find a database of suitable translations (Chap. 4), using these corpora to produce dictionaries and termbanks (Chap. 5), to build MT engines (Chap. 6) and to use them in other applications (Chap. 7).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

S. Sharoff et al., *Building and Using Comparable Corpora for Multilingual Natural Language Processing*, Synthesis Lectures on Human Language Technologies,
https://doi.org/10.1007/978-3-031-31384-4_1

1.1.1 Availability of Truly Parallel Data

The first limitation in the use of parallel corpora comes from the process of their production. For truly parallel corpora we need to collect texts that have been carefully translated by highly trained professional translators [8]. Many more people produce monolingual texts in their native languages in comparison to output of a small number of trained translators. Also there is an imbalance in the amount of translations produced for a relatively small number of major languages, primarily for the languages of the United Nations or the EU, while there are thousands of languages with very minor resources. Less resourced languages can still benefit from parallel resources. However, statistically speaking, language can be described as a large number of rare events in the sense that an individual word or expression is relatively infrequent, but the totality of rare events contributes to the mass probability of words and expressions in a text [9]. This creates problems of sparsity even for better-resourced languages. For example, the word *unicyclist* occurs 84 times in 2 billion words of a monolingual English ukWac corpus [10], i.e., about once per 23 million words in ukWac, with no examples of this word in the English parts of large publicly available parallel corpora such as Europarl [15] or the United Nations corpus [16].

Another constraint on corpus data concerns the availability of translation products, as it is easier to obtain translated texts produced by large public bodies, such as the European Parliament or the United Nations than many other kinds of translations. This bias leads to many word choices, which are specific with respect to the genres and topics available in such corpora. For example, there are 75 occurrences of the expression *strong voice* in Europarl, all of which are used in the sense of political authority, for example, *ensuring that smaller Member States retain a* ***strong voice*** *in the decision-making procedures*. At the same time, out of the 28 occurrences of this expression in the British National Corpus [17] 19 examples refer to the quality of human voices, for example, *She had a good,* ***strong voice****—an actor's voice*. When translating *strong voice* into other languages, the political authority metaphor needs to be explicitly unpacked, because the literal translation is not suitable. For example, when translating *strong voice* into Russian, the political sense is likely to use *reshitelno vystupatj* ('to express assertively') or *goryacho osuzhdat* ('to condemn vehemently') as opposed to the straightforward literal translation *gromkij golos* ('loud voice'). In the end, the mismatch between the domains and genres of parallel corpora and the target applications can lead to errors, which might be corrected by the use of less parallel resources.

1.1.2 Translationese in Parallel Data

Another kind of problems concerning the use of parallel data comes from a particular phenomenon known as translationese, namely a difference between features of translated texts and texts originally produced by native speakers [18]. Translationese is caused by factors inherent in the translation process, such as explicitation [19], i.e., the need to provide more information in the translated text for what remains implicit in the source text. For example,

translations tend to use more cohesive markers, such as *therefore, however, nevertheless*, in comparison to original texts by making the logical relations more explicit in translation [18]. Other factors leading to translationese are related to the temporal and cognitive pressures on the translation operations, as the translators need to complete their tasks in a short period of time. This leads such phenomena as (1) normalisation, i.e., the tendendency to re-use stock expressions of the target language even when the source text deviates from the norm, and (2) "shining-through" [20], i.e., the influence of the syntactic and lexical choices stemming from the source texts even when other choices are common in monolingually produced texts in the target languages. In the end it has been shown that translationese effects have statistical significance leading to the ability to build fairly accurate classifiers detecting texts translated by humans [21], even in an unsupervised fashion [22].

From the viewpoint of the translation direction, there is always a source text which needs to be translated into other languages, so only one text in a parallel corpus is primary, while other texts exhibits features of translationese. At the same time, when parallel corpora are used for MT applications, this directionality of translation is usually ignored, so that the Slovenian→ English MT built from the Europarl corpus uses slightly unnatural Slovenian texts exhitibing features of translationese as its source texts. These aspects call for the greater use of texts originally produced in the respective languages.

1.2 Levels of Comparability

These constraints on the availability and the use of parallel data led to numerous studies utilising less parallel resources. This research direction generally goes under the name of 'comparable corpora'. However, it is important to note that there is no clear dividing line between fully parallel and comparable corpora, as multilingual resources vary with respect to the degree of linking between documents in the two languages. Consider some examples of documents along the cline of comparability:

translations identifiable originally produced source texts and their translations.

- **truly parallel** true high quality translations, such as the proceedings of the European Parliament.
- **modified parallel** translations with some modifications to cater for the target audience. For example, language-specific descriptions of the Search dialogue box in the OpenOffice manual are translations from English with necessary modifications, such as searching for *New York* is replaced with searching for *Berlin* in the German version.
- **adaptations** translators exhibit freedom in rendering the source text, as it is the case with many of the fan-produced subtitles as in the OpenSubtitles corpus [23].

Table 1.1 Adena culture example

en	The Adena culture was a Pre-Columbian Native American culture that existed from 800 BC to 1 AD, in a time known as the Early Woodland period
de	Adena-Kultur ist die Bezeichnung für eine im mittleren Ohio-Tal ansässige prähistorische Indianerkultur. Sie lässt sich für die Zeit von etwa 1000 v. Chr. bis 200 n. Chr. nachweisen
fr	La Civilisation Adena était une culture pré-colombienne amérindienne ayant existé de l'an 1 000 à l'an 200 avant J.-C., durant l'ère connue sous le nom de Période sylvicole
ja	アデナ文化（Adena）は、アメリカ合国オハイオ州を中心に1000B.C. から元前後にえた文化。アデナ文化は、初期（ないし前期）ウッドランド期 （Early Woodlannd Period） の文化として位置付けられ、アデナ文化の出によって、後のホプウェル文化をはじめとするウッドランド文化の先となるウッドランド式土器や丘墓、トウモロコシ耕のすべてがでそろった。
ru	Культура Адена - доколумбовая индейская археологическая культура, существовавшая в период 1000-200 г. до н. э., в период, известный как ранний Будлендский период.

strongly comparable closely related texts produced in several languages.

Wikipedia entries entries on exactly the same topic are linked across languages via iWiki links, see Table 1.1, with individual entries varying in the amount of information.

news items very specific events are covered by numerous news agencies in various languages. Often the same agency reports the same story in various languages without fully relying on their translation, see the BBC News in English and Spanish.

weakly comparable similar texts which cannot be directly linked to each other across languages, while still being in the same domain and genre, for example:

- texts in the same narrow subject domain and genre, but describing different events, e.g. parliament debates on health care from the German Bundestag, the British House of Commons and the French parliament;
- texts within the same broader domain and genre, but varying in subdomains and specific genres, e.g. parallel queries in the renewable energy domain mostly returning wind energy research articles in English vs solar panel producers in Russian [24].

unrelated collections of unrelated texts, which are nevertheless collected using comparable methods from comparable sources. For instance, this concerns the use of random snapshots of the Web for Chinese, English, German and Russian [25] or the use of filtered Common Crawl data [26] with the assumption that different cultures use the Web for broadly similar purposes.