

SpringerBriefs in Statistics

More information about this series at <http://www.springer.com/series/8921>

George Tambouratzis · Marina Vassiliou
Sokratis Sofianopoulos

Machine Translation with Minimal Reliance on Parallel Resources

 Springer

George Tambouratzis
Institute for Language and
Speech Processing
Athens
Greece

Sokratis Sofianopoulos
Institute for Language and
Speech Processing
Athens
Greece

Marina Vassiliou
Institute for Language and
Speech Processing
Athens
Greece

ISSN 2191-544X
SpringerBriefs in Statistics
ISBN 978-3-319-63105-9
DOI 10.1007/978-3-319-63107-3

ISSN 2191-5458 (electronic)
ISBN 978-3-319-63107-3 (eBook)

Library of Congress Control Number: 2017947698

© The Author(s) 2017

This work is subject to copyright. All rights are reserved 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, express 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.

Printed on acid-free paper

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

Acknowledgements

The work presented in this book has originated from the PRESEMT project, which has comprised in total six partners, namely ILSP (Institute for Language and Speech Processing/Athena R.C.), GFAI (Gesellschaft zur Förderung der Angewandten Informationsforschung e.V.), NTNU (Norges Teknisk-Naturvitenskapelige Universitet), ICCS (Institute of Communication and Computer Systems), MU (Masaryk University) and LCL (Lexical Computing Ltd.).

The concept of the PRESEMT methodology was conceived within the Machine Translation Department of ILSP, in a collaborative effort during early 2009. The novelty of the concept is that it attempts to circumvent the requirement for specialised resources and tools so as to support the creation of MT systems for diverse language pairs without constraints. The authors wish to acknowledge within this process the contribution of other members of the department. In addition, we would like to note the contribution of the late Prof. George Carayannis, who was serving as the Director of ILSP and had a key role when starting the work to define the PRESEMT project. His presence is still felt in ILSP.

Though the work described in this book represents the work carried out at ILSP, the authors wish to acknowledge the substantial input of other partners in the creation of the final prototype, in terms of both the different modules and resources. Also, the authors wish to acknowledge the contribution of the late Adam Kilgarriff, Founder of LCL, with whom the Machine Translation Department of ILSP had a close association over more than 10 years. Adam Kilgarriff contributed to the implementation of PRESEMT in terms of assembling language resources.

PRESEMT was selected for funding by the European Commission in September 2009, within FP7/Call4. This funding has been instrumental in performing the related work, and the authors wish to acknowledge the contribution of this financial support (and the support of European Commission project officers as well as external reviewers) in completing the work summarised in this book.

Contents

1 Preliminaries	1
1.1 Challenges in MT—Relevance to the European Environment	1
1.2 A Brief Review of MT Development History	3
1.3 Advantages and Disadvantages of Main MT Paradigms	4
1.4 The PRESEMT Methodology in a Nutshell	7
1.5 Closing Note on Implementation	9
References	9
2 Implementation	11
2.1 Introduction: Summary of the Approach	11
2.2 Linguistic Resources: Data and Existing Linguistic Tools	13
2.2.1 External Processing Tools	13
2.2.2 Lemma-Based Bilingual Dictionary	15
2.2.3 The Parallel Corpus	15
2.2.4 The TL Monolingual Corpus	17
2.3 Processing the Parallel Corpus	19
2.3.1 Phrase Aligner Module	19
2.3.2 Phrasing Model Generation	24
2.4 Creating a Language Model for the Target Language	26
References	27
3 Main Translation Process	29
3.1 Introduction	29
3.2 Translation Phase One: Structure Selection	30
3.2.1 The Dynamic Programming Algorithm	32
3.2.2 Example of How Structure Selection Works	34
3.3 Phase Two: Translation Equivalent Selection	35
3.3.1 Applying the Language Model to the Task	37
3.3.2 Example of How TES Works	39
References	40

4	Assessing PRESEMT	43
4.1	Evaluation Dataset	44
4.2	Objective Evaluation Metrics	44
4.3	System Evaluation	45
4.3.1	Evaluation Objectives	45
4.3.2	Evaluation Results.	46
4.3.3	Expanding the Comparison	47
4.3.4	Experimenting with Further Data	47
4.4	Comparing PRESEMT to Other MT Systems.	49
4.5	Conclusions	52
	References.	53
5	Expanding the System	55
5.1	Preparing the System for New Language Pairs.	56
5.2	Examining Language-Pair-Specific Issues.	57
5.2.1	Agreement Within a Nominal Phrase	58
5.2.2	Case Mismatches.	58
5.2.3	The Null-Subject Parameter	58
5.2.4	Word Order.	59
5.3	Notes on Implementation	60
5.4	Conclusions	60
	References.	61
6	Extensions to the PRESEMT Methodology	63
6.1	Splitting SL Sentences into Phrases More Accurately.	63
6.1.1	Design and Implementation of TEM.	64
6.1.2	Experimental Evaluation	67
6.1.3	Conclusions.	68
6.2	Combining Language Models of Different Granularity	69
6.2.1	Extracting the N-Gram Models	71
6.2.2	Experimental Results.	72
6.2.3	Discussion.	74
	References.	74
7	Conclusions and Future Work	77
7.1	Review of the Effectiveness of the PRESEMT Methodology	77
7.2	Likely Avenues for Improvements in Translation Quality	78
7.2.1	Automatic Enrichment of Dictionary.	79
7.2.2	Design and Implementation of TEM.	79
7.2.3	Grouping of Tokens and PoS Tags into Related Classes	80
7.2.4	Revision of the Structure Selection Translation Phase.	81

7.2.5	Improving the Alignment of Words/Phrases	82
7.2.6	Augmenting the TL Language Model to Cover Supra-Phrasal Segments	82
7.2.7	A Closing Evaluation of Translation Accuracy	83
	References.	84
Glossary	87