Statistical Machine Translation

The field of machine translation has recently been energized by the emergence of statistical techniques, which have brought the dream of automatic language translation closer to reality. This class-tested textbook, authored by an active researcher in the field, provides a gentle and accessible introduction to the latest methods and enables the reader to build machine translation systems for any language pair.

It provides the necessary grounding in linguistics and probabilities, and covers the major models for machine translation: word-based, phrase-based, and tree-based, as well as machine translation evaluation, language modeling, discriminative training and advanced methods to integrate linguistic annotation. The book reports on the latest research and outstanding challenges, and enables novices as well as experienced researchers to make contributions to the field. It is ideal for students at undergraduate and graduate level, or for any reader interested in the latest developments in machine translation.

PHILIPP KOEHN is a lecturer in the School of Informatics at the University of Edinburgh. He is the scientific coordinator of the European EuroMatrix project and is also involved in research funded by DARPA in the USA. He has also collaborated with leading companies in the field, such as Systran and Asia Online. He implemented the widely used decoder Pharaoh, and is leading the development of the open source machine translation toolkit Moses.

Statistical Machine Translation

Philipp Koehn
School of Informatics
University of Edinburgh

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521874151

First published 2010
5th printing 2014

Printed in the United Kingdom by Print on Demand, World Wide

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-87415-1 Hardback

Additional resources for this publication at www.statmt.org

For Trishann and Phianna

Contents

Preface

Over the last few centuries, machines have taken on many human tasks, and lately, with the advent of digital computers, even tasks that were thought to require thinking and intelligence. Translating between languages is one of these tasks, a task for which even humans require special training.

Machine translation has a long history, but over the last decade or two, its evolution has taken on a new direction – a direction that is mirrored in other subfields of natural language processing. This new direction is grounded in the premise that language is so rich and complex that it could never be fully analyzed and distilled into a set of rules, which are then encoded into a computer program. Instead, the new direction is to develop a machine that discovers the rules of translation automatically from a large corpus of translated text, by pairing the input and output of the translation process, and learning from the statistics over the data.

Statistical machine translation has gained tremendous momentum, both in the research community and in the commercial sector. About one thousand academic papers have been published on the subject, about half of them in the past three years alone. At the same time, statistical machine translation systems have found their way to the marketplace, ranging from the first purely statistical machine translation company, Language Weaver, to the free online systems of Google and Microsoft.

This book introduces the major methods in statistical machine translation. It includes much of the recent research, but extended discussions are confined to established methodologies. Its focus is a thorough introduction to the field for students, researchers, and other interested parties, and less a reference book. Nevertheless, most of the recent research is cited in the *Further Readings* section at the end of each chapter.

I started this book about four years ago, and throughout its writing I have been supported by friends, family, and my wife Trishann, for which I am very grateful. I would not have been able to write this book without the ideas and opinions of the researchers who guide me, in this field: Kevin Knight, Franz Och, and Michael Collins. I am

also indebted to colleagues for providing comments, finding mistakes, and making valuable suggestions, most notably Eleftherios Avramidis, Steven Clark, Chris Dyer, Adam Lopez, Roland Kuhn, Paola Merlo, Miles Osborne, and Ashish Venugopal. This work would also not have been accomplished without the support of the members of the statistical machine translation group at the University of Edinburgh: Miles Osborne, Phil Blunson, Trevor Cohn, Barry Haddow, Adam Lopez, Abhishek Arun, Michael Auli, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Loïc Dugast, Hieu Hoang, David Talbot, and Josh Schroeder.

Edinburgh, October 17, 2008.