

# Statistical Machine Translation

**Philipp Koehn**

School of Informatics  
University of Edinburgh



# Contents

<i>Preface</i>	<i>page xi</i>
<b>I Foundations</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Overview	4
1.2 History of Machine Translation	14
1.3 Applications	20
1.4 Available Resources	23
1.5 Summary	26
<b>2 Words, Sentences, Corpora</b>	<b>33</b>
2.1 Words	33
2.2 Sentences	45
2.3 Corpora	53
2.4 Summary	57
<b>3 Probability Theory</b>	<b>63</b>
3.1 Estimating Probability Distributions	63
3.2 Calculating Probability Distributions	67
3.3 Properties of Probability Distributions	71
3.4 Summary	75
<b>II Core Methods</b>	<b>79</b>
<b>4 Word-Based Models</b>	<b>81</b>
4.1 Machine Translation by Translating Words	81
4.2 Learning Lexical Translation Models	87
4.3 Ensuring Fluent Output	94
4.4 Higher IBM Models	96
4.5 Word Alignment	113
4.6 Summary	118

<b>5</b>	<b>Phrase-Based Models</b>	<b>127</b>
5.1	Standard Model	127
5.2	Learning a Phrase Translation Table	130
5.3	Extensions to the Translation Model	136
5.4	Extensions to the Reordering Model	142
5.5	EM Training of Phrase-Based Models	145
5.6	Summary	148
<b>6</b>	<b>Decoding</b>	<b>155</b>
6.1	Translation Process	156
6.2	Beam Search	158
6.3	Future Cost Estimation	167
6.4	Other Decoding Algorithms	172
6.5	Summary	176
<b>7</b>	<b>Language Models</b>	<b>181</b>
7.1	N-Gram Language Models	182
7.2	Count Smoothing	188
7.3	Interpolation and Back-off	196
7.4	Managing the Size of the Model	204
7.5	Summary	212
<b>8</b>	<b>Evaluation</b>	<b>217</b>
8.1	Manual Evaluation	218
8.2	Automatic Evaluation	222
8.3	Hypothesis Testing	232
8.4	Task-Oriented Evaluation	237
8.5	Summary	240
<b>III</b>	<b>Advanced Topics</b>	<b>247</b>
<b>9</b>	<b>Discriminative Training</b>	<b>249</b>
9.1	Finding Candidate Translations	250
9.2	Principles of Discriminative Methods	255
9.3	Parameter Tuning	263
9.4	Large-Scale Discriminative Training	272
9.5	Posterior Methods and System Combination	278
9.6	Summary	283
<b>10</b>	<b>Integrating Linguistic Information</b>	<b>289</b>
10.1	Transliteration	291
10.2	Morphology	296
10.3	Syntactic Restructuring	302

10.4 Syntactic Features	310
10.5 Factored Translation Models	314
10.6 Summary	320
<b>11 Tree-Based Models</b>	<b>331</b>
11.1 Synchronous Grammars	331
11.2 Learning Synchronous Grammars	337
11.3 Decoding by Parsing	346
11.4 Summary	363
<i>Bibliography</i>	371
<i>Author Index</i>	416
<i>Index</i>	427

# 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.