Book Review

Statistical Machine Translation

Philipp Koehn
(University of Edinburgh)

Cambridge University Press, 2010, xii+433 pp; ISBN 978-0-521-87415-1, $60.00

Reviewed by
Colin Cherry
National Research Council Canada

Statistical Machine Translation provides a comprehensive and clear introduction to the most prominent techniques employed in the field of the same name (SMT). This textbook is aimed at students or researchers interested in a thorough entry-point to the field, and it does an excellent job of providing basic understanding for each of the many pieces of a statistical translation system. I consider this book to be an essential addition to any advanced undergraduate course or graduate course on SMT.

The book is divided into three parts: *Foundations*, *Core Methods*, and *Advanced Topics*. *Foundations* (75 pages) covers an introduction to translation, working with text, and probability theory. *Core Methods* (170 pages) covers the main components of a standard phrase-based SMT system. *Advanced Topics* (125 pages) covers discriminative training and linguistics in SMT, including an in-depth discussion of syntactic SMT. The text as a whole assumes a certain familiarity with natural language processing; though the *Foundations* section provides an effort to fill in the gaps, the book's focus is decidedly translation. As such, students unfamiliar with NLP may sometimes need to consult a general NLP text.

The book aims to provide a thorough introduction to each component of a statistical translation system, and it definitely succeeds in doing so. Supplementing this core material for each chapter is a highly inclusive *Further Reading* section. These sections provide brief narratives highlighting many relevant papers and alternative techniques for each topic addressed in the chapter. I suspect many readers will find these literature pointers to be quite valuable, from students wishing to dive deeper, to experienced SMT researchers wishing to get started in a new sub-field. Each chapter also closes with a short list of exercises. Many of these are very challenging (accurately indicated by a star-rating system), and involve getting your hands dirty with tools downloaded from the Web. The usefulness of these exercises will depend largely on the instructor's tastes; I view them as a bonus rather than a core feature of the book.

1. Chapters 1–3: Foundations

The first three chapters provide foundational knowledge for the rest of the book. *Introduction* provides an overview of the book and a brief history of machine translation, along with a discussion of applications and an expansive list of resources. The overview's structure takes the form of a summary of each chapter. This structure provides an effective preview of what will be covered and in what order, but it does not focus on typical introduction material; for example, there is no one place set aside to convince the reader that SMT is a good idea, or to introduce concisely the main philosophies behind the field. The history section is enjoyable, and I was glad to see a cautionary

note regarding machine translation's history of high hopes and disappointments. The applications section provides an excellent overview of where SMT sees actual use, and helps the reader understand why translations do not always need to be prefect.

Words, Sentences, Corpora provides a whirlwind tour of NLP basics, briefly touching on a broad set of topics including Zipf's law, parts-of-speech, morphology, and a number of grammar formalisms. To give an idea of just how brief coverage can be, the section on grammar covers four formalisms in five pages. Nonetheless, these descriptions should be helpful when the concepts re-appear later in the book. This chapter closes with a discussion of parallel corpora and sentence alignment. As these are central to the business of SMT, I feel they might have been better placed in a translation-focused chapter.

Probability Theory covers the basic statistics needed to understand the ideas throughout the book. This chapter is clear, and provides strong intuitions on important issues such as conditional probability. There is a surprisingly large emphasis on binomial and normal distributions, considering SMT's heavy reliance on categorical distributions; however, these are needed to discuss significance testing and some language modeling techniques covered later.

2. Chapters 4–8: Core Methods

The next five chapters provide detailed descriptions of each of the major components of a phrase-based SMT system. *Word-based Methods* discusses the five IBM translation models, with a brief detour to discuss the noisy channel model that motivates the IBM approach. This chapter is best taken as a complement to Brown et al. (1993) and Kevin Knight's (1999) tutorial on the same subject, rather than a replacement. It provides strong intuitions on what each IBM model covers and how each model works, including the clearest descriptions I have seen of IBM:3–5. However, it does sometimes make them seem a little mysterious. For example, there is no attempt to explain why IBM:1 always arrives at a global maximum, or to generalize when one can apply the mathematical simplification that reduces IBM:1's exponential sum over products to a polynomial product over sums. One glaring omission from this chapter is a discussion of the alignment HMM (Vogel, Ney, and Tillmann 1996). This elegant model is widely used and widely extended, and I had expected to see it covered in detail.

Chapters 5 and 6 on *Phrase-based Models* and *Decoding* cover the major algorithms in the popular phrase-based SMT paradigm. They are clear and fairly complete; this book could easily serve as an effective reference for these topics. *Phrase-based Models* motivates the use of phrases, and then covers phrase extraction along with the calculation of phrase features, such as lexical weighting and lexicalized re-ordering models. This chapter also marks the beginning of a careful dance, where log-linear models are introduced without having yet covered SMT evaluation or discriminative training. These topics are covered in Chapters 8 and 9, respectively. This division of modeling and training is a reasonable strategy, given the amount of material required to understand the full pipeline, but a student may need some extra guidance to understand the complete picture. The *Decoding* chapter focuses on stack decoding, and it is extremely well-written, with great explanations of search and pruning strategies. Alternative decoding formalisms, such as A* or finite-state decoding, are given short but effective summaries. This chapter makes phrasal SMT decoding feel easy.

The next chapter covers *Language Models*. Considering that this topic is given in-depth coverage in other NLP texts, I was surprised to see it covered quite thoroughly here as well. This chapter covers a number of smoothing techniques, as well as some

practical tips for handling large models. As usual, the exposition is exceptionally clear, and each new method's advantages are demonstrated with predicted counts or perplexity scores on Europarl data, which I found to be very useful. For many SMT courses, this chapter will be sufficient to stand alone as both an introduction and a reference for language modeling.

Finally, the *Core Methods* section closes with a discussion of *Evaluation*. This chapter discusses human evaluation, motivates automatic evaluation, and then covers the major contenders: word error rate, BLEU, and METEOR. The discussion of BLEU's shortcomings is very even-handed, perhaps a little pessimistic, and acknowledges all of the major concerns regarding the metric.

3. Chapters 9–11: Advanced Topics

The final three chapters cover advanced topics, which include recent or not universally adopted advances. So at this point, one might expect that all of the major components of a baseline phrase-based SMT system have been covered, but the final piece of the puzzle does not come until *Discriminative Training*, which includes a discussion of minimum error rate training (MERT) for the log-linear models introduced in Chapter 5. This chapter also covers *n*-best list extraction, *n*-best list re-ranking, and posterior methods such as Minimum Bayes Risk Decoding. It also devotes a surprisingly large amount of time to large-scale discriminative training, where thousands of parameter values can be learned. There is a lot of ground to cover here; consequently, much of the material will need to be supplemented with research papers or other texts if the instructor wants to cover any one topic in depth. The sections covering the learning methods used in parameter tuning (maximum entropy, MERT) did not feel as clear as the rest of the book. I suspect that a newcomer to the field will require some guidance to pick out the essential parts.

Chapter 10 is on *Integrating Linguistic Information*, which is kind of a grab bag, covering linguistic pre-processing, syntactic features, and factored translation models. The pre-processing discussion includes transliteration, morphological normalization, compound splitting, and even syntactically motivated re-ordering of the input sentence. The syntactic features section mostly covers *n*-best list re-ranking as done in the Smorgasbord paper (Och et al. 2004). Each of these topics is well motivated, and the text provides a clear description of a prominent, recent solution.

Finally, the book closes with *Tree-based Models*. This chapter covers a lot of ground: first describing synchronous context-free grammars, and then describing both formally syntactic hierarchical grammars and linguistically syntactic synchronous-tree-substitution grammars in terms of this common formalism. This is a very nicely presented chapter. It draws a lot of interesting connections between formalisms; for example, tree-to-tree rule extraction and tree-to-string rule extraction are presented as simple constraints on hierarchical phrase extraction. The description of chart parsing for decoding is also very clear, and it draws many useful analogies to the material presented earlier for phrasal decoding. I get the impression that many insights gained while adding syntactic SMT into the Moses translation system have found their way into this chapter.

4. Summary

This book's existence indicates that the field of SMT has reached a point of maturity where it makes sense to discuss core and foundational techniques. This book provides

a clear and comprehensive introduction to word, phrase, and tree-based translation modeling, along with the decoding, training, and evaluation algorithms that make these models work. The text's stated goal is to provide a thorough introduction, but I would also recommend it as an effective reference for anyone interested in writing their own SMT decoder, be it phrasal or syntactic. Most importantly, this book makes the prospect of teaching a course devoted to SMT much less daunting, and it should provide a valuable resource to researchers or students looking to teach themselves.

References

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312.

Knight, Kevin 1999. A statistical MT tutorial workbook. Available at: `http://www.isi.edu/~knight/`.

Och, Franz Josef, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenki Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zheng Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168, Boston, MA.

Vogel, Stephan, Hermann Ney, and Cristoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings, 16th International Conference on Computational Linguistics (COLING)*, pages 836–841, Copenhagen.

Colin Cherry is a research officer at the National Research Council Canada. His research interests include structure prediction and induction, with application to parsing, morphology, pronunciation, and machine translation. Cherry's address is NRC Institute for Information Technology, 1200 Montreal Road, M50:C-318, Ottawa, Ontario, Canada K1A 0R6; e-mail: `Colin.Cherry@nrc-cnrc.gc.ca`; URL: `https://sites.google.com/site/colinacherry/`.