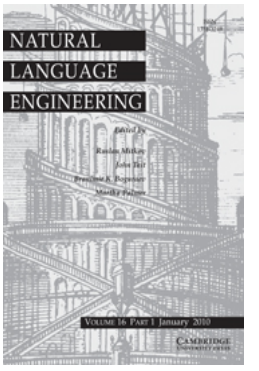
Learning Machine Translation



Learning Machine Translation, edited by Cyril Goutte, Nicola Cancedda, Marc Dymetman and George Foster. MIT Press, 2009.

Miles Osborne

Natural Language Engineering / Volume 16 / Issue 01 / January 2010, pp 99 - 100
DOI: 10.1017/S1351324909990222, Published online: 28 January 2010

Book Reviews

Learning Machine Translation, edited by Cyril Goutte, Nicola Cancedda, Marc Dymetman and George Foster. MIT Press, 2009.
doi:10.1017/S1351324909990222

Machine translation (MT) is a hot topic: at major conferences such as ACL, it often attracts 1 in 4 submissions. This shouldn't really be a surprise, since MT is such a hard problem and the payoffs are large. This importance has been noticed by machine learning researchers and the book "Learning Machine Translation" (LMT) arose from a workshop at the Neural Information Processing Systems Conference (NIPS) in December 2006. LTM contains papers originally presented at NIPS, along with original contributions. As such it offers value over-and-above the original workshop.

The book is divided into two major sections: one on background material and one on MT itself. The book starts with an excellent overview of MT, as seen in 2006. Not surprisingly, many of the topics covered in this chapter are still very much at the forefront of research –phrase based models, syntax, language modelling. Some predicted important directions (such as using kernels) have not yet reached maturity, and a number of directions, such as Bayesian approaches are not covered at all. The connection with speech processing is missed however, and the machine learning section is surprisingly thin. This chapter is likely to be useful for both machine learning researchers coming to the field, as well as new language researchers.

The rest of the first section contains a mixed bag of chapters. One chapter ("Mining Patents for Parallel Corpora") – which addresses the problem of getting more data for a particular domain – is rather specialised and of limited applicability. The next one ("Automatic Construction of Multilingual Name Dictionaries") is rather light from an MT perspective and seems like filler material. The chapter on Named Entity Translation contains solid machine learning, but again, the machine translation angle is brief. The chapter afterwards does address a core machine translation problem: word alignments. This chapter uses relatively naive machine learning approaches and doesn't present a clear message for translation; in particular, the approach they advocate always seems to hurt performance. A chapter on kernel methods for translation is nice and contains techniques that are tailored for natural language. One problem with this chapter (and in my experience, this is a common failing amongst machine learning researchers dabbling in language) is that the experimentation is modest. It seems here as if the authors care more about the techniques than the problem itself.

The second section of the book ("Machine Translation") contains chapters that are of more obvious interest to typical target readers. "Towards purely discriminative training for Tree-Structured Translation Models" presents a technique for applying large-scale discriminative training methods for translation. Since this paper was published, significant progress has been made in applying such techniques to MT.

The following chapter on reranking methods for MT presents simple techniques for improving performance using large amounts of data. One interesting lesson here is that often simple machine learning techniques yield gains and more complex approaches fail to deliver. This comment is supported by the chapter which follows, which presents an involved application of kernel methods for translation. The price to pay for using expensive methods is that they do not scale well, and this is reflected in the small-scale experiments. The next chapter ("Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction") presents a nice, larger-scale approach which clearly shows how MT can be viewed in terms of lexical selection and reordering. Afterwards follows a chapter which starts on discriminative phrase selection but leads onto a description of a set of non-standard evaluation metrics. The final two chapters are strong and deal with semi-supervised learning for MT and also system combination.

In summary, the book is a nice summary of the field as it was in 2006. Inevitably there has been much progress since then, so the book inevitably will become dated. Core truths remains however and a careful reading of the book will show that generally, progress in fields as mature and hard as MT are slow and unspectacular. Good machine learning is central, but so is an understanding of the problem itself.

Miles Osborne
Institute for Communicating and Collaborative Systems
The University of Edinburgh
Edinburgh, Scotland, UK

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. ISBN-13 978-0-521-86571-5, xxi + 482 pages.
doi:10.1017/S1351324909005129

When I was a student starting out in the field of information retrieval (IR) in the late 1980s, there was a general feeling that the IR community was lacking a definitively great IR book. Van Rijsbergen's 1979 *Information Retrieval* was a thoroughly researched document, which featured one of the first derivations of the probabilistic model of IR, but it was common to encounter readers who struggled to translate the book's ideas into actual IR systems. Salton's 1983, *Introduction to Modern Information Retrieval* had the practicalities excellently covered, but its author appeared to be less enthused about the coming theoretical models that are now an integral part of modern IR. For many years IR courses used both. Their dominance was eventually usurped in 1999 by *Modern Information Retrieval* edited by Baeza-Yates and Ribeiro-Neto. I reviewed it at the time and my feeling was that it was the best of the books on offer. Unfortunately, its outstanding chapters were balanced by a few middling ones and the odd important error; however, this book went onto be the standard that most undergraduate and postgraduate students turn to for information on IR.