Book Reviews

Bitext Alignment

Jörg Tiedemann
(Uppsala University)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 14), 2011, 153 pp; paperbound, ISBN 978-1-60845-510-2, $45.00; e-book, ISBN 978-1-60815-511-9, $30.00 or by subscription

Reviewed by
Michel Simard
National Research Council Canada

Bitext alignment techniques are at the heart of the revolution that swept through the field of machine translation (MT) two decades ago. From the early attempts of pioneers like Kay and Röscheisen (1993), to the recent success of [*enter name of your current favorite alignment method or researcher*], the growth of translation alignment is so tightly intertwined with that of statistical machine translation that it is actually difficult (and most likely futile) to try to establish which was most instrumental to which. From the beginning, researchers have been attracted by the alignment problem, not only because of its essential link to MT and the numerous other applications that could be derived, but also because bitext alignment seems like a "neat" problem: one around which a researcher can easily wrap his or her head until a clean solution emerges.

Of course, this neatness is mostly illusory, as many researchers eventually found out, and all these "possible" links in publicly available reference alignments are witnesses to the fact that translation alignment is not a perfectly well defined problem, and that explaining the subtle ways in which identical concepts are rendered across languages by drawing lines between words is an oversimplification, to say the least.

In spite of this, Tiedemann's book is a timely addition to the natural language processing literature. After going through the motivations in Chapter 1, he introduces the basic concepts and terminology in Chapter 2, discusses various alignment types, models and search procedures, and presents the fundamentals of alignment evaluation. Of particular interest is the discussion on the crucial role played by the segmentation of the text into the units on which the alignment will operate. The process of collecting and structuring parallel corpora is briefly outlined in Chapter 3.

In Chapter 4, Tiedemann gets to the heart of the matter with sentence-level alignment. Following a more-or-less historical line, he first covers approaches based on surface features (sentence lengths, alignment type), à la Gale and Church (1993), then methods relying on lexical resources, before looking at combined and resource-specific methods.

Chapter 5 covers word-alignment techniques. As one would expect, it makes up a sizeable portion of the book (about a third). Half of this chapter is devoted to generative translation models (essentially, the IBM models), the other half to various discriminative models. Finally, Chapter 6 rapidly surveys phrase and tree alignment models.

Overall, Tiedemann's book covers a lot of ground, possibly a bit too much. This is a vast field, which has seen the publication of hundreds (thousands?) of publications over the last 20 years or so. The author clearly knows his stuff, and manages to structure its presentation in a logical and intuitive manner. The exposition's clarity sometimes suffers, however, from the author's obvious desire not to miss anything. Many research avenues are just alluded to, and even fundamental notions are sometimes presented in a sketchy manner. For instance, the topic that clearly gets the most elaborate presentation is the IBM models, with 15 pages; yet it is unlikely that a reader new to this field will manage to extract more than a general intuition about them.

It is also worth noting that the book probably requires more from the reader than just "familiarity with basic concepts" of machine learning. The means and methods of machine learning have become so ubiquitous in computational linguistics that we tend to forget how fundamental they are: the cycle of training and testing, feature engineering, the debates of generative versus discriminative modeling, supervised versus semi-supervised versus unsupervised learning, and so forth. All of this is part of our daily routine and has taken over our way of viewing the field.

All this is not to say that this is not a useful book. It is a well-structured, well-written overview of the state-of-the-art in text-translation alignment. It reads like a roadmap of the work accomplished more than a true travel guide. As such, it is rich with pointers to the many variations on methods and approaches to the problem. It will most likely be of interest to the bi-curious among us: graduate students and researchers who, although already familiar with computational linguistics, may feel attracted to the other text.

References

Gale, W. A. and K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Kay, M. and M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1): 121–142.

This book review was edited by Pierre Isabelle.

Michel Simard obtained a Ph.D. in computer science from the Université de Montréal in 2003, but has been actively involved in natural language processing research since 1986: He was part of the machine-aided translation team of CITI, a research laboratory of the Canadian Department of Industry; was a founding member of the RALI laboratory at the Université de Montréal; and was a postdoctoral researcher with the Machine Learning group at Xerox Research Centre Europe (Grenoble, France). Now a senior research officer at the National Research Council Canada, his work focuses mainly on machine translation and machine-assisted translation. Simard's e-mail address is michel.simard@nrc-cnrc-gc.ca.