Empirical Methods for Exploiting Parallel Texts

I. Dan Melamed
(New York University)

Cambridge, MA: MIT Press, 2001,
xi+195 pp; hardbound, ISBN
0-262-13380-6, $32.95, £22.95

Reviewed by
Ted Pedersen
University of Minnesota, Duluth

Parallel translations of written texts have long been useful tools for human students of language and have begun to serve as an intriguing source of data for corpus-based approaches to natural language processing. A source text and its translation can be viewed as a coarse map between the two languages, and an industrious student or clever computer program may wish to refine that mapping so that it shows which sentences, phrases, and words are translations of one another.

Humans are very adept at finding such relations in parallel text. This is true even when one or both of the languages is unfamiliar, as can be seen in a simple but convincing exercise by Knight (1997). Although there was considerable early success in automatically identifying sentences in parallel text that are translations of each other (e.g., Brown, Lai, and Mercer 1991, Gale and Church 1993), a variety of challenging problems has emerged since that time.

Empirical Methods for Exploiting Parallel Texts is a revision of author I. Dan Melamed's 1998 Ph.D. dissertation (University of Pennsylvania) and succeeds in capturing the range of problems inherent in parallel text. It presents a variety of techniques for finding translation equivalents and demonstrates that once these are available, they can be used to align text segments, detect omissions in translations, identify noncompositional compounds, and discriminate among word senses.

The book is arranged in three parts, the guiding organizational principle of which is the distinction between tokens and types. (A token represents each occurrence of a linguistic entity in a text, whereas a type consists of every occurrence of identical tokens.) The author casts his own work in terms of pattern recognition techniques that acquire information about specific tokens and statistical learning methods that induce generalized models of word types on the basis of token data.

Part I consists of three chapters that focus on tokens and pattern recognition. Chapter 2 presents an algorithm that finds a token-by-token mapping of a parallel text in which each token in the source text is aligned with its translation in the target text. This algorithm is presented in terms of pattern recognition concepts such as signal generation, filtering, and search.

The signals generated are candidate token alignments, and these come from cognates identified in the parallel text and optionally from a user-supplied seed lexicon of word translations. High-frequency words create noise that is filtered by means of a localized search procedure that allows the algorithm to consider the parallel text piece by piece as it performs token alignment. The next two chapters (3 and 4) show that once such a mapping is available, it can be used to find segment alignments in parallel text and to detect portions of text that have been erroneously omitted from a translation.

Part II is a two-chapter interlude that touches on both tokens and types. Chapter 5 is a discussion of the issues involved in counting co-occurrences in parallel text. Although Melamed's conclusions may seem fairly intuitive, he shows that previous work in translation modeling has often been based on less-than-optimal counting schemes. In Chapter 6, Melamed describes the creation of a manually word-aligned version of the Bible in French and English. This is a nontechnical chapter that captures the difficulties of manual alignment in general and in dealing with the Bible in particular.

The three chapters in Part III consider word types and statistical learning techniques. Chapter 7 develops three statistical models of translation that represent unordered word-by-word translation. These models rely on the observation that a source word will most often translate into a single target word (as opposed to multiple words) or may have no translation equivalent at all. Parameter estimation schemes are developed that capture these properties, and empirical results show that the inclusion of parameters developed according to these schemes steadily improves the translation models, all of which significantly outperform the historically important IBM Model 1 (Brown et al. 1993).

Melamed argues that word-by-word translation models can also be used to identify linguistic entities that are normally not translated word for word. In Chapter 8 he shows how noncompositional compounds such as *beat a dead horse* can be discovered in parallel text by means of such a model. In Chapter 9, an unsupervised word sense discrimination algorithm is introduced that is evaluated by treating the different discovered senses of a word as distinct types. The incorporation of these additional sense distinctions as types is shown to improve the quality of a translation model.

Several of the chapters in this book have appeared in preliminary form in this journal (Chapters 2, 3, and 7) and conference proceedings (Chapters 4 and 8). As a result, the individual chapters are well polished and clearly the product of careful reviewing and rewriting. Since this material retains its original style and form, each chapter is relatively self-contained, and readers may wish to approach this book as a collection of related papers rather than as a work that must be read from start to finish.

Readers not familiar with corpus-based methods, pattern recognition, and statistical learning may struggle at times. There is not a great deal of preliminary or background material presented in each chapter before moving into more advanced concepts. This is particularly true of Chapters 2 and 7, which present the token alignment algorithm and translation model development. This should not discourage the novice from approaching this book, however. Since each chapter can be read independently, it is possible to appreciate the related applications described in Chapters 3, 4, 8, and 9 without fully understanding all of the technical details in Chapters 2 and 7. The author also includes sufficient discussion of related work throughout to provide a general picture of the field.

To conclude, this is a useful book for anyone interested in parallel text. It offers enough improvements and clarifications over previously published material that it will appeal to a reader already familiar with the author's work. It also has a wide scope and strikes a reasonable balance between theory and application, so newcomers will get a sense of the possibilities and problems associated with parallel text. These readers may wish to supplement this book with pattern recognition and statistical-learning readings, as found in Duda and Hart (1973), for example. Finally, in an age in which academic texts tend to be expensive and poorly printed, this book is a notable exception, as the production quality is high and the price is reasonable.

References

Brown, Peter F., Stephen Della Pietra, Vincent Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Brown, Peter F., Jennifer Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, CA.

Duda, Richard O. and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley.

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

Knight, Kevin. 1997. Automating knowledge acquisition for machine translation. *AI Magazine*, 18(4):81–96.

Ted Pedersen is an Assistant Professor of Computer Science at the University of Minnesota, Duluth. His research focuses on word sense disambiguation. He is the recipient of a National Science Foundation CAREER award. Pedersen's address is: Department of Computer Science, University of Minnesota, Duluth, MN 55812; e-mail: tpederse@d.umn.edu.