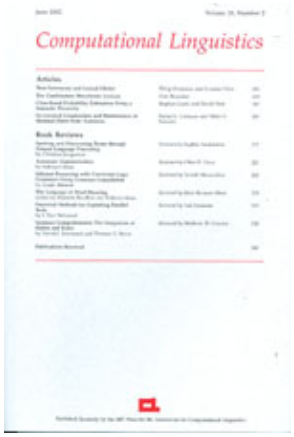
Empirical Methods for Exploiting Parallel Texts I. Dan Melamed (New York University) Cambridge, MA: MIT Press, 2001, xi+195 pp; hardbound, ISBN 0-262-13380-6, $32.95, £22.95



Quarterly (March, June, September, December)

160pp. per issue

6 3/4 x 10

Founded: 1974

2018 Impact Factor: 1.319

2018 Google Scholar h5-index: 32

ISSN: 0891-2017

E-ISSN: 1530-9312

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

Ted Pedersen

Posted Online March 30, 2006

https://doi.org/10.1162/coli.2002.28.2.235

Computational Linguistics
Volume 28 | Issue 2 | June 2002
p.235-237

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 f
of language and have begun to serve as an intriguing source of da
approaches to natural language processing. A source text and its
viewed as a coarse map between the two languages, and an ind
clever computer program may wish to refine that mapping so t
sentences, phrases, and words are translations of one another.

Humans are very adept at finding such relations in parallel te
when one or both of the languages is unfamiliar, as can be se
convincing exercise by Knight (1997). Although there was consid
in automatically identifying sentences in parallel text that are transl
(e.g., Brown, Lai, and Mercer 1991, Gale and Church 1993), a va
problems has emerged since that time.

Empirical Methods for Exploiting Parallel Texts is a revision of autho
1998 Ph.D. dissertation (University of Pennsylvania) and succee
range of problems inherent in parallel text. It presents a variety of
ing translation equivalents and demonstrates that once these are
be used to align text segments, detect omissions in translations, id
tional compounds, and discriminate among word senses.

The book is arranged in three parts, the guiding organizationa
is the distinction between tokens and types. (A token represents
a linguistic entity in a text, whereas a type consists of every occ
tokens.) The author casts his own work in terms of pattern recogni
acquire information about specific tokens and statistical learning n
generalized models of word types on the basis of token data.

Part I consists of three chapters that focus on tokens and p
Chapter 2 presents an algorithm that finds a token-by-token ma
text in which each token in the source text is aligned with its tran
text. This algorithm is presented in terms of pattern recognition con
generation, filtering, and search.

The signals generated are candidate token alignments, and th
nates identified in the parallel text and optionally from a user-su
of word translations. High-frequency words create noise that is fi
a localized search procedure that allows the algorithm to consid
piece by piece as it performs token alignment. The next two chap
that once such a mapping is available, it can be used to find segr
parallel text and to detect portions of text that have been erroneou
translation.

Forthcoming