March 1993 Volume 19, Number 1

Computational Linguistics

Special Issue on Using Large Corpora: I

Articles

Squibs and Discussions

Book Reviews

Published Quarterly by the MIT Press for the Association for Computational Linguistics

Computational Linguistics is abstracted and/or indexed in the following publications: *Annual Review of Information Science and Technology, Computer Abstracts, Computing Reviews, Current Contents: Social and Behavioral Sciences, Knowledge Engineering Review, Language and Language Behavior Abstracts, Linguistic Abstracts, Linguistic Bibliography, Science Abstracts* (INSPEC Section C), and *Social Sciences Citation Index.*

Volume 14 and subsequent volumes of *Computational Linguistics* are available on microfilm through University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106, USA.

Individual Subscriptions to *Computational Linguistics* are available only with membership in the Association for Computational Linguistics (ACL). Annual Membership Dues are as follows: *regular,* $30.00; *student* — taking a full-time program of courses and not earning a regular income, $20.00; *retired,* $20.00; *joint* — for couples receiving one subscription, $40.00. A membership application together with an order form for back issues and conference proceedings is included at the end of this issue. Send membership applications and address changes to Dr. Donald E. Walker (ACL), Bellcore, 445 South Street, MRE 2A379, Morristown, NJ 07960, USA; acl@bellcore.com. However, note that it is also possible to pay by credit card (both by e-mail and by postal mail) and through banks in Switzerland and Japan.

Institutional Subscriptions are $80.00. Outside the U.S. add $14.00 postage and handling. Canadian subscribers also add 7% GST. Send institutional subscription orders and address changes to MIT Press Journals, 55 Hayward Street, Cambridge, MA 02142-1399, USA. To be honored free of charge, claims for missing copies must be made immediately upon receipt of the next published issue. Institutions should order back issues (before 1988) and all proceedings from the ACL at the address above.

Preface

Susan Armstrong-Warwick
ISSCO, University of Geneva
54, Route des Acacias, 1227 Geneva,
Switzerland

Introduction

The papers in this two-volume special issue on "using large corpora" bear witness to promising developments in computational linguistics. Empirical and statistical methods offer new means of organizing data and attaining insights into our use of language. From a practical point of view, they provide a basis for progress in the performance of NLP systems. This collection of papers represents the diversity of work on these new methods and their application to a range of problems. They offer partial solutions to the key issues of acquisition, coverage, robustness, and extensibility. What they all have in common is that the models they employ arise from the use of large collections of texts for training, organization of data, and evaluating success.

Background to This Issue

When the idea first arose to publish a special issue of CL on using large corpora, the topic was not generally considered to be part of mainstream CL, in spite of an active community working in this field. The rapidly increasing number of papers devoted to this topic were not yet circulating in the public domain, and large text collections were not easily obtainable. One commonly held view was that this was a rather marginal activity, limited to specific applications and only pursued in a few centers, and the few papers that were in wide circulation were often viewed with skepticism. On the other hand, a growing number of researchers had recognized the potential of exploiting large corpora and were making considerable progress in developing new methods and demonstrating practical results. The first motivation for devoting a special issue to this topic was thus to bring the numerous and diverse studies based on corpora to the entire CL community.

Since the call was first published, the corpus-based approach has established itself as an important development in CL. The interest in this new direction has become apparent in the number of papers, invited talks, and tutorial sessions now commonplace at NLP conferences and the number of workshops and meetings devoted entirely to this topic.

What is it that has brought about this rapid growth of interest in corpus-based NLP? For some, it is simply a rediscovery of empirical and statistical methods popular in the 1950s. Machine translation, for example, was at that time viewed as a 'mere' decoding problem, but computing resources were far from adequate for processing the data according to this model. The technological advances in computer power has certainly favored the reintroduction of this approach, as has the growing availability of large-scale textual resources in machine-readable form.[1]

[1] The ACL/Data Collection Initiative, the European Collection Initiative (sponsored by the ACL,

More important, perhaps, is the growing frustration of trying to use standard rule-based methods to account for more than a well-chosen fragment of text, regardless of the application. The data extracted from large corpora have demonstrated that language use is more flexible and complex than that which most rule-based systems have up to present tried to account for. The relative lack of practical results at a time when industrial concerns are looking to the CL community to demonstrate progress toward useful applications has also contributed to the growing interest in new methods. And finally, the success rate demonstrated in the speech community offers hope for similar progress in NLP.

Themes and Topics

The new data-oriented methods offer potential solutions to key problems:

- acquisition: automatically identifying and coding all of the necessary information
- coverage: accounting for all of the phenomena in a given domain, collection of texts, application, etc.
- robustness: accommodating 'real data' that may be corrupt or not accounted for in the model
- extensibility: applying the model and data to a new domain, a new set of texts, a new problem, etc.

While these problems are not new, the exploitation of corpora offers new directions in which solutions may be sought. Aspects of these topics are addressed in the papers in various ways: learning word patterns (i.e., subcategorization, collocations, translation) from their appearance in the texts, improving robustness, and adaptability to new domains. A theme to be found in a number of the papers that indirectly touches on the problems listed above is which methods are appropriate for the task at hand. A topic related to this is the question of which types of information we can characterize using statistical methods. In terms of traditional categories these two volumes contain papers on lexical issues (including morphology and syntactic and semantic relations), issues in syntax (such as incorporating probability measures in parsing and discussing attachment problems), and translation topics (from building tools for alignment to models for fully automatic MT).

A last topic to be addressed with regard to this issue and the use of data-oriented methods in general is the numerous contrasts and controversies arising in the discussions. The debate has been characterized in terms of 'statistics-based vs. rule-based,' 'empirical vs. rationalist,' 'Chomsky-inspired vs. Shannon-inspired,' etc.[2] To those not directly engaged in this debate, however, such matters may seem of secondary interest. Many of the papers that directly address this topic in these two volumes demonstrate how the two approaches can be combined, rather than arguing for one in opposition to the other. As stated at the beginning, my purpose in putting together this special issue has been to identify new methods and describe the potential results that the use of large corpora offers us.

ELSNET, and other European institutions), and the creation of the Linguistic Data Consortium have helped to make a growing collection of data widely available.

[2] The introductory article by K. Church and B. Mercer, solicited for this volume, puts some of this discussion into a historical context and provides a personal perspective on some of the underlying issues.