Title: ACL Data Collection Initiative

URL: https://en.wikipedia.org/wiki/ACL_Data_Collection_Initiative

PageID: 79553235

Categories: Category:Association for Computational Linguistics, Category:Computational linguistics, Category:Datasets, Category:Datasets in machine learning, Category:Natural language processing, Category:Speech recognition

---

The ACL Data Collection Initiative (ACL/DCI) was a project established in 1989 by the Association for Computational Linguistics (ACL) to create and distribute large text and speech corpora for computational linguistics research. The initiative aimed to address the growing need for substantial text databases that could support research in areas such as natural language processing , speech recognition , and computational linguistics . By 1993, the initiative's activities had effectively ceased, with its functions and datasets absorbed by the Linguistic Data Consortium (LDC), which was founded in 1992. [1]

Objectives

The ACL/DCI had several key objectives:

To acquire a large and diverse text corpus from various sources

To transform the collected texts into a common format based on the Standard Generalized Markup Language (SGML)

To make the corpus available for scientific research at low cost with minimal restrictions

To provide a common database that would allow researchers to replicate or extend published results

To reduce duplication of effort among researchers in obtaining and preparing text data

These objectives were designed to address the growing demand for very large amounts of text arising from applications in recognition and analysis of text and speech. Its core objective was to "oversee the acquisition and preparation of a large text corpus to be made available for scientific research at cost and without royalties". [2]

History

By the late 1980s, researchers in computational linguistics and speech recognition faced a significant problem: the lack of large-scale, accessible text corpora for developing statistical models and testing algorithms. Existing generally available text databases were too small to meet the needs of developing applications in text and speech recognition. The initiative was formed to meet this need by collecting, standardizing, and distributing large quantities of text data with minimal restrictions for scientific research. As stated by Liberman (1990), "research workers have been severely hampered by the lack of appropriate materials, and specially by the lack of a large enough body of text on which published results can be replicated or extended by others." [2]

The ACL/DCI committee was established in February 1989. The committee included members from academic and industrial research laboratories in the United States and Europe. [3]

The initiative was chaired by Mark Liberman from the University of Pennsylvania (formerly of AT&T; Bell Laboratories). Other committee members included representatives from organizations such as Bellcore , IBM T.J. Watson Research Center , Cambridge University , Virginia Polytechnic Institute & State University , Northeastern University , University of Pennsylvania , SRI International , MCC , Xerox PARC , ISSCO , and University of Pisa . [3]

The project operated initially without dedicated funding, relying on volunteer efforts from committee members and their affiliated institutions. Key supporters included AT&T; Bell Labs, Bellcore, IBM,

Xerox, and the University of Pennsylvania, which allowed the use of their computing facilities for ACL/DCI-related work. [2]

Previously running on volunteer effort pro bono , in 1991, it obtained funding from General Electric and the National Science Foundation (IRI-9113530). [4]

Data

As of 1990, the ACL/DCI had collected hundreds of millions of words of diverse text. The collection included: [2] [3]

Wall Street Journal articles (25 to 50 million words);

Canadian Hansard (parliamentary records) in parallel English and French versions: cleaned-up English Hansard donated by the IBM alignment models group (100 million words), and original Bilingual Hansard (from a different time period) obtained directly (200 million words).

Collins English Dictionary (1979 edition), both as fulltext (3 million words) and as various "database" versions, constructed using "typographers' tape" donated by Collins, which were computer tapes containing the structured digital data used to typeset and print the 1979 edition of the dictionary;

Emails from ARPANET newsletters for the ACM Special Interest Group on Information Retrieval Forum (IRLIST) and AIList Digest issues distributed over the ARPANET (AILIST) (5 million words), both collected by Edward A. Fox at VIPSU ;

Articles on networking (2 million words);

U.S. Department of Agriculture Extension Service Fact Sheets (>1 million words);

200,000 scientific abstracts of about 1,500 words each from the Department of Energy (25 million words);

Archives of the Challenger Investigation Commission , including transcripts of depositions and hearings (2.5 million words);

Books from the Library of America , including works by Mark Twain , Eugene O'Neill , Ralph Waldo Emerson , Herman Melville , W.E.B. DuBois , Willa Cather , and Benjamin Franklin (130 books, 20 million words);

Public domain books like the King James Bible , Tristram Shandy , The Federalist Papers ;

Several million words of transcribed radiologists ' reports, donated by Francis Ganong at Kurzweil Applied Intelligence Inc (about 5 million words);

The Child Language Data Exchange corpus of child language acquisition transcripts; [5]

U.S. Department of Justice Justice Retrieval and Inquiry System (JURIS) materials; [6]

The Swiss Civil Code in parallel German, French and Italian;

Economic reports from the Union Bank of Switzerland , in parallel English, German, French and Italian;

About 12K words of administrative policy manuals and 14K words of administrative memos, contributed by Geoff Pullum of U.C.S.C. ;

Material from various ACM journals and the ACL journal Computational Linguistics ;

The CSLI publications series: 50-100 reports (8K words each) and 5-10 books (80K words each).

The initiative started with North American English text but expanded to include Canadian French and planned to include Japanese, Chinese, and other Asian languages. [2]

At least 5 million words from the collection were tagged under the Penn Treebank project, and those tags were distributed by DCI as well. [2] [3] [7]

After DCI was absorbed by the LDC, the datasets were curated under LDC. [8]

Format

The ACL/DCI corpus was coded in a standard form based on SGML (Standard Generalized Markup Language , ISO 8879), [2] consistent with the recommendations of the Text Encoding Initiative (TEI), of which the DCI was an affiliated project. The TEI was a joint project of the ACL, the Association for Computers and the Humanities , and the Association for Literary and Linguistic Computing , aiming to provide a common interchange format for literary and linguistic data.

The initiative planned to add annotations reflecting consensually approved linguistic features like part of speech and various aspects of syntactic and semantic structure over time. [2]

Examples

As an example of the use of ACL/DCI, consider the Wall Street Journal (WSJ) corpus for speech recognition research. The WSJ corpus was used as the basis for the DARPA Spoken Language System (SLS) [9] community's Continuous Speech Recognition (CSR) Corpus. [10] The WSJ corpus became a standard benchmark for evaluating speech recognition systems and has been used in numerous research papers.

The WSJ CSR Corpus provided DARPA with its first general-purpose English, large vocabulary, natural language, high perplexity corpus containing speech (400 hours) and text (47 million words) during 1987–89. The text corpus was 313 MB in size. [10]

The text was preprocessed to remove ambiguity in the word sequence that a reader might choose, ensuring that the unread text used to train language models was representative of the spoken test material. The preprocessing included converting numbers into orthographics , expanding abbreviations , resolving apostrophes and quotation marks , and marking punctuation . [10]

As another example, the Yarowsky algorithm used bitext data from DCI to train a simple word-sense disambiguation model that was competitive with advanced models trained on smaller datasets. [11]

Distribution

Materials from the ACL/DCI collection were distributed to research groups on a non-commercial basis. By 1990, about 25 research groups and individual researchers had received tapes containing various portions of the collected material. [2]

To obtain the data, researchers had to sign an agreement not to redistribute the data or make direct commercial use of it. However, commercial application of "analytical materials" derived from the text, such as statistical tables or grammar rules, was explicitly permitted. [2]

The initiative first distributed data via 12-inch reels of 9-track tape , then via CD-ROMs. Each such tape could contain 30 million words compressed via the Lempel-Ziv algorithms . [2] The first CD-ROM distribution was in 1991, funded by Dragon Systems Inc . It contained Collins English Dictionary, WSJ, scientific abstracts provided by the U.S. Department of Energy, and the Penn Treebank. [4]

See also

Linguistic Data Consortium

Penn Treebank

Text Encoding Initiative

Computational linguistics

Natural language processing

Speech recognition

References