Parallel Text Collections at Linguistic Data Consortium

Xiaoyi Ma
Linguistic Data Consortium
3615 Market St. Suite 200
Philadelphia, PA 19104, USA
xma@ldc.upenn.edu

Abstract The Linguistic Data Consortium (LDC) is an open consortium of universities, companies and government research laboratories that creates, collects and distributes speech and text databases, lexicons, and other resources for research, teaching and technology development. This paper describes past and current work on the creation of parallel text corpora, and reviews existing and upcoming collections at LDC.

1 Introduction

There is increasing interest in computer-based linguistic technologies, including speech recognition and understanding, optical and pen-based character recognition, text retrieval and understanding, machine translation, and the use of these methodologies in computer assisted language learning. In each area, we have useful present-day systems and realistic expectations of progress. However, because human language is so complex and information-rich, computer programs for processing it require enormous amounts of linguistic data - speech, text, lexicons, and grammars - to be robust and effective. Such databases are expensive to create, document and maintain. Even the largest companies have difficulty acquiring enough of this data to satisfy their research and development needs. Researchers at smaller companies and in universities risk being frozen out of the process almost entirely. For pre-competitive research, shared resources also provide benefits that closely-held or proprietary resources do not. Shared resources permit replication of published results, support fair comparison of alternative algorithms or systems, and permit the research community to benefit from corrections and additions provided by individual users.

Until recently, most linguistic resources were not generally available for use by interested researchers. Because of concern for proprietary rights, or because of the additional burdens of electronic publication (which include preparation of a clean and well-documented copy, securing of clear legal rights and drafting of necessary legal agreements, and subsequent support), most of the linguistic databases prepared by individual researchers have either remained within a single laboratory, or have been given to some researchers but refused to others.
A few notable examples over the years have demonstrated the value of shared resources, but until recently, these have been the exceptions rather than the rule. For example, the Brown University text corpus has been used by many researchers, to the point of being adopted as a generally-available test corpus for evaluating statistical language models of English. The importance of shared data for evaluation of speech technology was shown by the TI-46 and TI DIGITS databases, produced at Texas Instruments in the early 1980's, and distributed by the National Institute of Standards and Technology (NIST) starting in 1982 and 1986 respectively. The U.S. Defense Department's Advanced Research Projects Agency (ARPA) began using a "common task" methodology in its speech research program in 1986, creating a series of shared databases for algorithm development and evaluation. This approach led to rapid progress in speech recognition, and has since been applied to research in message understanding, document retrieval, speech understanding, and machine translation.

Building on these successes, the Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes more than 260 companies, universities, and government agencies. Since its foundation, the LDC has delivered data to over 600 unique organizations including non-member institutions.

2 Parallel Text Collections at LDC

In last ten years, large parallel corpora have proven to be extremely useful for research in multilingual natural language processing, such as statistical machine translation [Brown 1990] [Melamed 1998], cross-lingual information retrieval [Davis & Dunning 1995] [Landauer & Littman 1990] [Oard 1997], lexical acquisition [Gale & Church 1991a] [Melamed 1997a]. Large parallel corpus are also very useful in language education, especially second language learning.

The LDC has been involved in the collection and creation of parallel text corpora over the past years. The current catalog includes the Canadian Hansard, United Nations Proceedings, European Corpus Initiative Multilingual Corpus.

2.1 Canadian Hansard

The Canadian Hansard Corpus consists of parallel texts in English and Canadian French, drawn from official records of the proceedings of the Canadian Parliament. While the content is limited to legislative discourse, it spans a broad assortment of topics and the stylistic range includes spontaneous discussion and written correspondence along with legislative propositions and prepared speeches.

LDC's Hansard corpus was assembled from two distinct secondary sources. Material from one time period of parliamentary proceedings was acquired through the IBM T. J. Watson Research Center, while material from another period was acquired through Bell Communications Research Inc. (Bellcore). The combined collection covers a time span from the mid-1970's through 1988, with no apparent duplication between the two data sources.
Aside from covering different time periods, the two archives have different organization and have undergone different amounts and kinds of processing in the preparation of the corpus. In addition, the Bellcore set itself comprises two distinct types of data - one appears to be the main parliamentary proceedings (similar in nature to the IBM set), while the other consists of transcripts from committee hearings. The three sets have been kept distinct in this publication and each is described in greater detail in the corpus documentation.

The three data sets have a number of features in common:

- They are rendered using the 8-bit ISO-Latin1 character-encoding standard.
- They use a minimal amount of SGML tagging to identify sentences or paragraphs.
- All sets are organized using a parallel file structure, in which the content of a given English text file is matched by the content of a corresponding French text file.

There are also several differences between the three sets:

- The IBM collection is presented as a sequence of parallel sentences (there are nearly 2.87 million parallel sentence pairs in the set).
- The Bellcore data are presented as sequences of paragraphs.
- The Bellcore main-session data is accompanied by mapping files that provide computed paragraph alignments and word-token correspondences; no additional alignment data are provided for the Bellcore committee texts (and none are needed for the IBM sentences).

2.2 United Nation Parallel Text:

This corpus was provided to the LDC by the United Nations, for use in research on machine translation technology. The documents come from the Office of Conference Services at the UN in New York and are drawn from archives that span the period between 1988 and 1993.

This UN Parallel Text corpus contains English, French and Spanish archives, with data from each language stored on a separate disc in the set. Care has been taken to arrange the document files in a parallel directory structure for each language, so that corresponding translations of a document are found directly by means of the directory paths and file names.

All parallel files in this corpus are English-based: for every file on the English disc, there will be a corresponding file on either the French or Spanish disc, or both. Tables are included on all discs to assist in determining which parallels are present. Due to the nature and organization of UN translation services and the original electronic text archives, the process of finding and sorting out parallel documents yielded a numerous gaps, with many files in each language having no parallel in other languages.

In preparing the text for publication, LDC applied a fully-compliant SGML format (Standard Generalized Markup Language) with a working DTD (Document Type Definition) provided on each disc. For those who do not need SGML markup, a simple script is included that can be used to filter out the SGML-specific material and leave only the plain text. The character set used is the 8-bit ISO 8859-1 Latin1, in which accented letters and some other non-ASCII characters occupy the upper 128 entries of the character table.

2.3 European Corpus Initiative Multilingual Corpus

The ECI is a volunteer effort, sponsored by the Association for Computational Linguistics (European Chapter), carried out at the Human Communication Research Centre, University of Edinburgh (HCRC) and Institute Dalle Molle pour les etudes semantique et cognitives, University of Geneva (ISSCO), with modest additional financial support from the European Network in Language and Speech (ELSNET) and the Network for European Reference Corpora (NERC).

3 Upcoming Release – Chinese – English Parallel Text:

LDC is now working on a major release of Chinese – English parallel text. Immediately upon Hong Kong's return to China on July 1st, 1997, both Chinese and English became the official languages of Hong Kong Special Administration Region. The Hong Kong government publishes almost all of its official publications in both languages. As a result, the website of Hong Kong government has become a valuable resource for Natural Language Processing research community.

Over last year, LDC has been collecting, cleaning, and aligning parallel texts from the Hong Kong Special Administration Region (HKSAR) government website. The corpora are divided into three parts with respect to their content: The Laws of HKSAR, press releases and news items of HKSAR and the Hong Kong Hansard. The corpora will be sentence aligned and will retain the original Big5 encoding. The Laws of HKSAR includes all the laws as of January 1999. 238,721 sentences in each language. The corpus is 61 Meg bytes. The Hong Kong Special Administration Region publishes press releases and news items online everyday. LDC has collected the data from July 1st, 1997 to the present. The overall collection is about 60 Meg bytes and grows at a rate of about 6 Meg bytes per month.
The Hong Kong Hansards consist of parallel texts drawn from official records of the proceedings of the Hong Kong Parliament. The corpus contains all the weekly meeting records of Hong Kong parliament since 1995. The overall size of the collection is about 70 Meg bytes.

4 Future Works

Many worldwide web sites are bilingual or multilingual. As the size of Internet continues to grow, the internet is becoming a gold mine of parallel text.

Over the last year, LDC developed a program, Bilingual Internet Text Search (BITS) [Ma1999], which crawls the World Wide Web in search of parallel text in specific language pairs. Our experiments showed that there are large quantities of parallel text on the Internet and that BITS technology is very successful in finding it.

Used BITS, we plan to harvest of parallel text over the WWW; initially to increase our Chinese – English collection and subsequently to adopt new language pairs, such as Korean – English, Thai – English, Vietnamese – English, and Indonesian – English.

References:

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roosin, P. (1990). “A statistical approach to machine translation”. Computational Linguistics, 16(2), 79-85.
Church, K.W., & Mercer, R. (1993). “Introduction to the special issue on computational linguistics using large corpora”, Computational Linguistics, 19(1), 1-24.

Davis, M., & Dunning, T. (1995). “A TREC evaluation of query translation methods for multi-lingual text retrieval”. Fourth Text Retrieval Conference (TREC-4). NIST.

Dunning. T. (1994). “Statistical identification of language”. Computing Research Laboratory technical memo MCCS 94-273, New Mexico State University Las Cruces, New Mexico.

Gale, W. A., & Church, K. W. (1991a). “Identifying word correspondences in parallel texts”. Fourth DARPA Workshop on Speech and Natural language, Asilomar, California.

Gale, W. A., & Church, K. W. (1991b). “A program for aligning sentences in bilingual corpora”. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California.

Landauer, T. K., & Littman, M. L. (1990). “Fully automatic cross-language document retrieval using latent semantic indexing”. Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, pp.

Pages 31-38, UW Centre for the New OED and Text Research, Waterloo, Ontario.

LDC (1999). Linguistic Data Consortium (LDC) home page. http://www.ldc.upenn.edu/.

Ma, Xiaoyi (1999) “BITS: A Method for Bilingual Text Search over the World Wide Web”. The MT Summit VII, submitted.

Melamed, I. D. (1996). “A geometric approach to mapping bitext correspondence”. Conference on Empirical Methods in Natural Language Processing, Philadelphia, Pennsylvania.

Melamed, I. D. (1997). “Automatic discovery of noncompositional compounds in parallel data”. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), Brown University.

Melamed, I. D. (1998). “Word-to-word models of translation equivalence”. IRCS technical report #98-08, University of Pennsylvania.

Oard, D. W. (1997). “Cross-language text retrieval research in the USA. Third DELOS Workshop”. European Research Consortium for Informatics and Mathematics.

Resnik, P., & Melamed, I. D. (1997). “Semi-automatic acquisition of domain-specific translation lexicons”. Fifth Conference on Applied Natural Language Processing, Washington, D.C.

Resnik, P., Olsen, M. B., & Diab, M. (1998). “The Bible as a parallel corpus: Annotating the ‘Book of 2000 Tongues’”.