Resources for Chinese-English Cross-Language IR

Douglas W. Oard
College of Library and Information Services
University of Maryland
College Park, MD 20742

February 27, 1999

This work has been supported in part by DARPA contract N6600197C8540

Part I. Introduction

Specific evaluation criteria are identified in the sections that follow. A stoplight chart (with λ representing fully suitable, "ω" representing suitability limited in some way, and "μ" representing possibly unsuitable) has been used to summarize the assessment of each criterion for each resource in each language. Missing data is indicated by leaving the cell blank. The stoplight charts in the last three parts of this document can provide a basis for resource selection and for gap analysis to focus additional search for specific resources that could satisfy unmet requirements.

Part II. Document Preprocessing Resources

Section Definition. Resources for identifying and normalizing the terms contained in documents before indexing them.

Encoding, Character Set, and Language Identification

Definition. Resources for recognizing commonly encountered character sets and their encodings and for identifying the dominant language in which a document is written. Automatic segmentation of a single document into regions written in different languages is not generally performed by these tools. There are two commonly used character sets for traditional Chinese characters: GB and Unicode. GB can serve as its own encoding, or it can be encoded as HZ if it is desired to interleave it with 7-bit ASCII. In either case, it is possible to represent Chinese and English simultaneously since the GB character set does contain Roman characters. There are three commonly used character sets for simplified Chinese characters: Big-5 (which, with some extensions, is also known as "CP-950"), CNS 11643, and Unicode. Big-5 and CNS 11643 are normally encoded by simply interleaving it with 7-bit ASCII, for which there are provisions in the code space. Thus, both can represent Chinese and English simultaneously.

Evaluation Methodology. Encodings that can accommodate multiple character sets (e.g., ISO-2022 and the Extended Unix Code (EUC)) can provide unambiguous indications of the encoded character set, and such encodings are typically designed to be distinguishable from each other. Many character sets routinely serve as their own encoding (e.g., ASCII, GB, and Big-5), however, and such codes are rarely designed to be easily distinguishable from each other. This can be an important issue for systems that must process unrestricted text from a wide variety of sources, since *a priori* knowledge (perhaps inferred from the source or provided as HTML markup) or some automatic character set detection technique is needed. Kikui (1996) applied algorithms for converting several language-specific encodings to a common representation, applied an automatic language identification technique to each candidate, and selected the candidate that produced the best match between the hypothesized encoding and the identified language. Overall detection accuracy and a confusion matrix depicting the predominance of misclassifications between each possible encoding pair were presented. Reeder & Geisler (1998) are working on character set identification techniques to handle a greater number of languages, but evaluation results have not yet been published. Language identification accuracy depends on the similarity of the language pairs that might potentially be confused and on document length. The most common evaluation methodology is to report confusion matrices for a representative set of passage lengths. Except for certain language pairs, language identification accuracy typically asymptotically approaches 100% for longer documents, so evaluations generally focus on relatively short passages. Developers of commercial software rarely report the effectiveness on short passages, however, so that does not serve as a useful discriminator among commercially available systems.

Evaluation Criteria.

- Availability
  - λ Available now
  - μ Projected to become available in 1999
- Cost
  - λ Free, or available with a multiuser license for $100 or less
  - ω Available for sale on negotiated terms or at a fixed price that exceeds $100 for multiple users
- Format
  - λ Available in an easily readable digital format
  - ω Available in a suitable format, but requires extensive preprocessing
  - μ Available only in hardcopy, data entry and validation estimated at over 40 hours
- Coverage
  - λ All present languages in ISO-8859-1 and all common character sets for Chinese
  - ω At least one common character set for Chinese
- Rejection effectiveness
  - λ Provisions are included to explicitly reject unknown character sets.
  - ω A fairly large number of character sets can be recognized, and hence rejected.
  - μ No effective provisions are included for rejection of undesired character sets.

Name	Availability	Cost	Format	Coverage	Rejection Effectiveness
codeguess	λ	λ	λ	ω	ω
Intelliscope	λ	ω	λ		λ
Mitre	μ	λ	λ	λ	λ
Que	λ	ω	λ	ω	λ

codeguess

A character set guesser designed specifically for Chinese characters written by Erik Peterson

- Availability
  Available by HTTP from http://www.erols.com/eepeter/codeguess.html
- Cost
  The source code contains a statement that the software is free for noncommercial use but that a fee is required for commercial use of the software.
- Format
  Distributed as a PERL 5 source code.
- Coverage
  The software is designed to recognize only the character set, rather than the character set and language. The software can recognize GB, HZ, Big-5, and ASCII. CNS 11643 and ISO-8859-1 are not supported, although the author has stated an interest in supporting CNS 11643 in the future.
- Rejection effectiveness
  A confidence threshold is used to reject unknown languages, but the omission of ISO-8859-1 may compromise its effectiveness.

IntelliScope
The Lernout & Hauspie "IntelliScope Language Recognizer" includes character set recognition, character set conversion and language recognition, but only the languages handled are specified.

- Availability
  Presently available from Lernout & Hauspie N.V., a Belgian company, through their Language Technologies division. Intelliscope was originally developed by Inso Corp. Additional information is available at http://www.lhsl.com/tech/icm/retrieval/toolkit/lr.asp.
- Cost
  IntelliScope can be licensed for a fee, but the price is not stated in their advertising literature.
- Format
  Apparently distributed as precompiled binaries for Windows 95/NT and Solaris 2.3. Functions are accessed through an API.
- Coverage
  The character set coverage is not explicitly stated, but the advertising materials explicitly state that both traditional and simplified characters are recognized, and all present languages are recognized.
- Rejection effectiveness
  Intelliscope can distinguish 36 languages.

Mitre
Flo Reeder of Mitre is developing a character set and language identification tool.

- Availability
  The software is being developed under government contract, and release authority from the government contracting officer is required. Chinese is included within the set of languages to be supported, but a version of the software with Chinese recognition capabilities is not yet ready for distribution.
- Cost
  There is no charge for use of the software on government projects, but any required support may incur costs. Terms for commercial use would need to be discussed with the government contracting officer.
- Format
  Once authority to use the software is granted, the software and the associated training data can be obtained through FTP.
- Coverage
  Very comprehensive character set coverage for each language is planned.
- Rejection effectiveness
  The software will eventually distinguish multiple encodings of more than 30 languages.

Que
The Alis Qué system for the identification of language and character encoding.

- Availability
  Presently available as part of the Flores toolkit from Alis Technologies, Inc., a Canadian company. Described at http://www.alis.com/castil/silc/index.en.html.
- Cost
  Que is offered on a license fee basis that varies based on the estimated value added to the end-user product.
- Format
  Que is distributed as Windows 95/NT and Solaris 2.x binaries with a C/C++ API designed for use with gcc on Solaris and Visual C++ on Windows 95/NT.
- Coverage
  Que can recognize GB, HZ, Big-5, and all present languages in the ISO-8859-1 character set, but not CNS 11643.
- Rejection effectiveness
  Que can recognize 28 languages and 98 language-encoding pairs.

Character Set Conversion

Definition. Resources for mapping other character sets into Unicode.

Evaluation Methodology. The Unicode standard, now in version 2.1, has essentially been stable for the present languages and the languages of interest since it was merged with ISO 10646 in 1993 to produce version 1.1 o the standard. Unicode does, however, define alternative representations (known as "composed" and "decomposed") for some characters. Although standards-compliant Unicode applications are required to handle all representations correctly, normalization to a uniform representation is typically needed in information retrieval applications. A description of Unicode normalization issues can be found in McCallum (1993), and the present status of the standard with respect to normalization is described in Davis (1998).

Evaluation Criteria.

- Availability
  - λ Available now
  - μ Projected to become available in 1999
- Cost
  - λ Free, or available with a multiuser license for $100 or less
  - ω Available for sale on negotiated terms or at a fixed price that exceeds $100 for multiple users
- Format
  - λ Available in an easily readable digital format
  - ω Available in a suitable format, but requires extensive preprocessing
  - μ Available only in hardcopy, data entry and validation estimated at over 40 hours
- Character set coverage
  - λ ISO-8859-1 and all common character sets for Chinese
  - ω At least one common character set for Chinese
- Unicode normalization
  - λ Includes normalization functions
  - ω Generates Unicode, but does not provide explicit control over normalization
  - μ Generates a single code other than Unicode that could be converted to Unicode

Traditional and simplified characters occupy different portions of the Unicode code space. Because there is a many-to-one mapping from traditional to simplified characters, an irreversable normalization from traditional to simplified characters is required when a query expressed using one type of characters must be matched with documents written using the other type of characters. The CNS 11643 character set evolved from Big-5 and large portions of the two code sets are identical, so applying a Big-5 converter to CNS 11643 is a reasonable engineering solution when a CNS 11643 converter is not available if occasional errors can be tolerated.

Name	Availability	Cost	Format	Char Set Coverage	Unicode Normalization
Flores	λ	ω	λ	ω	ω
MUTT	λ		λ	ω	ω
Rosette	λ	ω	λ	ω	λ

Flores

The Alis Flores/Bantam "universal character set conversion engine."

- Availability
  Presently available as part of the Flores toolkit from Alis Technologies, Inc., a Canadian company. Described at http://www.alis.com/castil/flores/conversions.en.html.
- Cost
  Both the Flores toolkit and the Bantam library are offered on a license fee basis that varies based on the estimated value added to the end-user product.
- Format
  Available in the Flores toolkit as a C or C++ API, either as source code or precompiled for Windows or Solaris. The same capabilities are available as a Windows 95/NT DLL with a C++ API in the Alis Bantam Library.
- Character set coverage
  Converts bidirectionally between GB, HZ or Big-5 and Unicode. CNS 11643 is not handled as a separate character set from Big-5. Source code for a CNS 11643 to Big 5 converter is available in b5cns.tar.gz from http://www.ifcss.org/ftp-pub/software/unix/convert/.
- Unicode normalization
  No explicit control over Unicode normalization is apparent in the documentation.

MUTT

The Multilingual Unicode Toolkit (MUTT) is a set of Unicode tools for display and conversion of text that was developed at New Mexico State University.

- Availability
  MUTT is available at ftp://crl.nmsu.edu/pub/misc/. A description of the toolkit is available at http://crl.nmsu.edu/Research/Projects/oleada/mutt.html.
- Cost

- Format
  MUTT is distributed as precompiled binaries for Solaris 2.1. Installation of Tcl/Tk is required.
- Character set coverage
  Converts bidirectionally between GB or Big-5 and Unicode. CNS 11643 is not handled as a separate character set from Big-5, and the HZ encoding is not handled. Source code for a CNS 11643 to Big 5 converter is available in b5cns.tar.gz and for a HZ to GB converter is available in HZ-2.0.tar.gz. Both are available from http://www.ifcss.org/ftp-pub/software/unix/convert/.
- Unicode normalization
  MUTT provides no control over Unicode normalization.

Rosette

The Rosette C++ library for Unicode was developed by Basis Technology Corp.

- Availability

  Presently available as a commercial product from Basis Technology Corp. Uniconv, a full-featured precompiled standalone demonstration or Rosette for noncommercial use, is available at http://unicode.basistech.com/

- Cost

  Both Rosette and uniconv can be licensed for a fee, but the price is not stated in their advertising materials.

- Format

  Rosette is available as C++ source code. Uniconv is available in binary form, with versions for Windows 95/NT and for Solaris 2.5.

- Character set coverage

  Converts bidirectionally between GB, HZ, or Big-5 and Unicode. CNS 11643 is not handled as a separate character set from Big-5. Source code for a CNS 11643 to Big 5 converter is available in b5cns.tar.gz from http://www.ifcss.org/ftp-pub/software/unix/convert/.

- Unicode normalization

  Unicode normalization functions are included.

Segmentation and Compound Splitting

Definition. Resources for segmenting texts in languages such as Chinese that lack orthographic boundaries between words.

Evaluation Methodology. Segmentation and compound splitting are instances of the more general term selection problem. As an abstract task, term selection is inherently an ill-formed problem because no single level of granularity is universally appropriate. Information retrieval systems designed for English text often cope with this situation by retaining multiple levels of granularity (e.g., both multiword terms and constituent phrases). Compound splitting and segmentation have, however, been predominantly studied at the component level, and thus the commonly used evaluation methodology is to compare the postulated segment boundaries with a single "gold standard" segmentation that is produced by a native speaker of the language. The reported statistics are normally derived from the number of missed segment boundaries and the number of incorrectly postulated segment boundaries.

Evaluation Criteria.

- Availability
  - λ Available now
  - μ Projected to become available in 1999
- Cost
  - λ Free, or available with a multiuser license for $100 or less
  - ω Available for sale on negotiated terms or at a fixed price that exceeds $100 for multiple users
- Format
  - λ Available in an easily readable digital format
  - ω Available in a suitable format, but requires extensive preprocessing
  - μ Available only in hardcopy, data entry and validation estimated at over 40 hours
- Accuracy
  - λ Appears to be built using the best known techniques
  - ω Fails to exploit some known techniques that could improve performance
- Unicode compatibility
  - λ Works directly on Unicode representations
  - ω Works on character representations that could be generated from Unicode

Name	Availability	Cost	Format	Accuracy	Unicode Compatibility
Flores	λ	ω	λ		λ
ch_seg	λ		λ	λ	ω
segmenter	λ	λ	λ	ω	ω

Flores
The Alis Flores toolkit includes a "word extraction" capability that performs segmentation.

- Availability
  Presently available as part of the Flores toolkit from Alis Technologies, Inc., a Canadian company. Described at http://www.alis.com/castil/flores/faq.html
- Cost
  The Flores toolkit is offered on a license fee basis that varies based on the estimated value added to the end-user product.
- Format
  Available in the Flores toolkit as a C or C++ API, either as source code or precompiled for Windows or Solaris.
- Accuracy
- Unicode compatibility
  Flores is Unicode-based.

ch_seg
The ch_seg Chinese segmentation was developed by Lei Chen at New Mexico State University

- Availability
  Presently available by FTP from ftp://crl.nmsu.edu/pub/misc/.
- Cost
- Format
  The software is distributed as C source code.
- Accuracy
  The algorithm was developed for a Masters Thesis at one of the best computational linguistics laboratories.
- Unicode compatibility
  The software is designed to work with GB rather then Unicode.

segmenter
A program developed by Erik Peterson to perform segmentation on Chinese text.

- Availability
  Presently available by HTTP from http://www.erols.com/eepeter/segmenter.html
- Cost
  The software is freely available. It contains no statements either granting or restricting rights for commercial use.
- Format
  The software is distributed as PERL source code.
- Accuracy
  No accuracy figures are reported, and the description of the algorithm suggests that some useful sources of information are not yet exploited.
- Unicode compatibility
  The software is designed to work with GB rather than with Unicode.

Proper Name Resources

Definition. Lists of names for individuals, organizations, and geographic features in a language of interest, preferably with categories assigned to those names. Monolingual training corpora in which proper names are tagged in a way that could support machine learning algorithms for proper name recognition such as those described by Gallippi (1996) are also of interest.

Evaluation Methodology. The MUC-6 "named entity" evaluation methodology is described in DARPA (1995) and in Hirschman (1998), and the evolution of the MUC-6 task is described in Grishman & Sundheim (1996). Chinese, Japanese, English and Spanish evaluations have been conducted using the same methodology in what is known as the "multilingual Entity Task" (MET). MET-1 was described in the TIPSTER Phase II workshop (DARPA 1996), and MET-2 is described at ftp://ftp.muc.saic.com/pub/MET/participation/call-for-participation. The MUC-7 named entity evaluation methodology is essentially the same as that used for MUC-6 and both MET evaluations. It is described at http://www.muc.saic.com/scorer/Manual/manual.html. The task is essentially one of classification, so a single result set is computed. Recall and precision are then computed using a hand-scored evaluation corpus. A version of van Rijsbergen's F measure (van Rijsbergen 1979) in which recall and precision are weighted equally is typically reported as a single figure of merit for each participating system.

Evaluation Criteria.

- Availability
  - λ Available now
  - μ Projected to become available in 1999
- Cost
  - λ Free, or available with a multiuser license for $100 or less
  - ω Available for sale on negotiated terms or at a fixed price that exceeds $100 for multiple users
- Format
  - λ Available in an easily readable digital format
  - ω Available in a suitable format, but requires extensive preprocessing
  - μ Available only in hardcopy, data entry and validation estimated at over 40 hours
- Category coverage
  - λ Handles person names, organization names and location names
  - ω Handles at least one of those categories
- Domain coverage
  - λ Broad coverage of names expected to occur in general news and technical texts
  - ω Moderate coverage of names expected to occur in general news
  - μ Some potentially useful names
- Unicode compatibility
  - λ Encoded in Unicode
  - ω Encoded in a character representation that could be converted to Unicode

Name	Availability	Cost	Format	Category Coverage	Domain Coverage	Unicode Compatibility
cweb	λ	λ	λ	ω	μ	ω
MET	λ	λ	λ	λ	μ	ω

cweb
A web-accessible bilingual term list hosted at the National Chiao-Tung University in Taiwan that contains city, country and personal names in Chinese and English.

- Availability
  Presently available through HTTP from http://www.csie.nctu.edu.tw/center/cweb/.
- Cost
  The files are freely available. They contain no statements either granting or restricting rights for commercial use.
- Format
  Each file is available as text and as HTML.
- Category coverage
  There is a file for location names (country.txt, country.html) and a file for person names (name.txt, name.html)
- Domain coverage
  The place names file contains the names of 179 countries and their capitals and the person names file contains 459 popular given names in China, the UK, and the US.
- Unicode compatibility
  The files are encoded in Big-5.

MET
The training collection for the Multilingual Entity Tasks (MET-1 and MET-2) included Chinese training materials in which proper names were marked.

- Availability
  MET training data was distributed to participants by FTP. The procedures are described at ftp://ftp.muc.saic.com/pub/MET/participation/call-for-participation.
- Cost
  There was no charge for participation in MET, but the available materials do not detail any provision for providing the material to nonparticipants, nor whether there are any restrictions on the use of the material as a basis for derivative works.
- Format
  The MET training data was distributed using a password-protected FTP site.
- Category coverage
  Person, organization and location names were hand-tagged in the MET training corpus.
- Domain coverage
  The Chinese materials in MET-2 were drawn from the Xinhua news agency, the Peoples Daily newswire, and China Radio transcripts.
- Unicode compatibility
  A sample of the MET-2 evaluation corpus that appears to be in GB code is available at http://www.muc.saic.com/scorer/gui/Chi/texts.

Part IV. Resources for Mapping Terms Between Languages

Definition. Resources such as thesauri, ontologies, lexicons, terminology lists, and cognate matching rules that explicitly specify relationships between terms in Chinese and terms in English.

Evaluation Methodology. For resources that lack conceptual structure (as is the case for all Chinese resources listed below), the most salient factor is size. Melamed (1995, 1997) developed a fully automatic methodology for assessing the match between a translation lexicon and an unannotated evaluation corpus of parallel documents. The evaluation corpus is first automatically aligned at the sentence level using dynamic programming techniques. Translations that appear in the lexicon are scored as valid if an occurrence of a word in one language within a source-language sentence is matched in any position of the corresponding target-language sentence by any possible translation of that word. An alternate methodology based on manually annotated ground truth was developed for the MUC-6 and MUC-7 "template element" tasks. That task evaluated the ability of participating systems to recognize alternate forms for person and organization names in English text based on evidence from an individual document. Recall, precision and the F measure are computed over each name and each alias that could be extracted from every document (allowing duplicates if they are in different documents). In MUC, other template elements (e.g., person title and organization location) were also scored in the evaluation corpus, so published results on the template element task are confounded with tasks that are extraneous to term-term matching. Furthermore, the MUC template element evaluation methodology confounds entity name recognition with entity name matching. Source code for the MUC scoring software is available from ftp://ftp.muc.saic.com/pub/MUC/scorer/, MUC-6 evaluation material is available on the ACL/DCI disk, and the ground truth markup may be available from SAIC. The MUC-6 and MUC-7 template element evaluations have been conducted only in English, so similar evaluation resources may not be available in other languages. A simpler methodology was used by Knight & Graehl (1997) to evaluate back-transliteration. In Knight's back-transliteration experiment the goal was to select the English word from which a Japanese katakana transliteration had been generated. Accuracy figures were reported for 100 personal names selected from a bilingual dictionary. As formulated by Knight & Graehl, back-transliteration is a more challenging task than transliteration matching because a single correct English counterpart must be selected.

Evaluation Criteria.

- Availability
  - λ Available now
  - μ Projected to become available in 1999
- Cost
  - λ Free, or available with a multiuser license for $100 or less
  - ω Available for sale on negotiated terms or at a fixed price that exceeds $100 for multiple users
- Format
  - λ Available in an easily readable digital format
  - ω Available in a suitable format, but requires extensive preprocessing
  - μ Available only in hardcopy, data entry and validation estimated at over 40 hours
- Lexicon size
  - λ Large lexicon (over 100,000 unique roots or multiword terms in the language of interest)
  - ω Moderate-sized lexicon (10,000 to 100,000 unique roots or terms in language of interest)
  - μ Small lexicon (fewer than 10,000 unique word roots or terms in the language of interest)
- Domain coverage
  - λ General news and technical terminology
  - ω General news terminology
  - μ Potentially useful technical terminology
- Morphology
  - λ Complete inflectional morphology
  - ω Moderately robust inflectional morphology
  - μ Spotty or no coverage of inflectional morphology
- Accuracy
  - λ Hand constructed of hand verified
  - ω Automatically built from corpora
- Translation preference information
  - λ A rich set of domain-specific translation probabilities are provided
  - ω Either a translation preference order or a single preferred translation is provided
  - μ No translation preference information is provided
- Unicode compatibility
  - λ Encoded in Unicode
  - ω Encoded in a character representation that could be converted to Unicode

Name	Availability	Cost	Format	Lexicon Size	Domain Coverage
CEDICT	λ	ω	λ	ω	ω
cweb	λ	λ	λ	μ	λ
ecdict	λ		ω	ω	ω
eng-chi	λ	λ	λ	λ	ω
TwinBridge	λ	ω	ω	ω	

Name	Morphology	Accuracy	Translation Preference	Unicode Compatibility
CEDICT	μ	λ	μ	ω
cweb	ω	λ		ω
ecdict	μ	λ	λ	ω
eng-chi	μ	λ	λ	ω
TwinBridge	μ	λ		

CEDICT

Paul Denisowski's CEDICT Chinese-English dictionary.

- Availability
  Freely downloadable from http://www.mindspring.com/~paul_denisowski/cedict.html.
- Cost
  The README file states that noncommercial use is permitted, but that permission is required from the copyright holder for commercial use. No cost is stated.
- Format
  CEDICT is stored in the same format at Jim Breen's Japanese-English EDICT.
- Lexicon size
  In September. 1998 CEDICT contained 16,830 entries.
- Domain coverage
  It appears from the documentation that CEDICT presently consists mostly of general terminology.
- Morphology
  No morphology information is included with CEDICT.
- Accuracy
  The entries in CEDICT have been manually verified.
- Translation preference information
  It does not appear that translation preference information is encoded in CEDICT.
- Unicode compatibility
  CEDICT is available in both the GB and Big-5 character sets.

cweb

A unidirectional Chinese to English bilingual dictionary.

- Availability
  Freely downloadable using HTTP from http://www.csie.nctu.edu.tw/center/cweb/.
- Cost
  The dictionary is freely available, and there is no statement granting or restricting permission for commercial use.
- Format
  Each file contains one English word per line, with (possibly) several Chinese translations for each English word.
- Lexicon size
  From the file sizes, it appears that several thousand total words are present in the various lists.
- Domain coverage
  There are a few files for general terminology and 92 short files for technical terminology in a number of fields.
- Morphology
  The entries are grouped in a way that may provide useful information about morphology.
- Accuracy
  The bilingual term lists appear to have been constructed manually.
- Translation preference information
  It is not clear whether alternative translations are presented in preference order.
- Unicode compatibility
  The Chinese terms are encoded in Big-5.

ecdict

http://www.ok88.com/go/svc/ecdict.html

Version 2.12 of an online English to Chinese dictionary developed by Linda Ng.

- Availability
  The dictionary is provided by OK88 Bilingual Internet Services.
- Cost
  There is no information about commercial availability of the dictionary.
- Format
  It appears that the dictionary is available only as an online resource. A complete list of English words beginning with any letter can be displayed, so automated retrieval of every English to Chinese translation would be practical.
- Lexicon size
  The reported size of the dictionary is 12,054 entries.
- Domain coverage
  The dictionary contains only general terminology.
- Morphology
  Only root forms are present in the dictionary, and no morphology information is provided.
- Accuracy
  The dictionary appears to be constructed by hand.
- Translation preference information
  Only a single translation is provided for each English word.
- Unicode compatibility
  The Chinese characters are coded in the Big-5 character set.

eng-chi
A Chinese-English bilingual term list.

- Availability
  Available through HTTP at http://www.math.psu.edu/simpson/chinese/ChinText/b5/eng-chi.
- Cost
  The term list is freely available, and there is no statement granting or restricting permission for commercial use.
- Format
  The dictionary is available as a text file, with one translation equivalent word pair per line.
- Lexicon size
  There are 103,000 word pairs in the term list.
- Domain coverage
  The term list contains general terminology.
- Morphology
  The English words are root forms, and no information about morphology is provided.
- Accuracy
  The term list appears to have been hand constructed or hand validated.
- Translation preference information
  There is a single translation for each English word.
- Unicode compatibility
  The dictionary is coded in the Big-5 character set. A version that has been automatically converted to GB is available at http://www.math.psu.edu/simpson/chinese/ChinText/gb/eng-chi.

TwinBridge
The TwinBridge bidirectional English-Chinese Dictionary.

- Availability
  Presently available from The TwinBridge Software Corp. Additional information is available at http://www.twinbridge.com/.
- Cost
  $99 for the end-user version. No price for access to the dictionary through an API is stated.
- Format
  Available on CDROM as an end-user product with an integrated user interface designed for any language version of Windows 95. The dictionary is clearly designed for human use and it appears that extensive reformatting would be needed to produce a usable cross-language lexicon. It is not clear whether an API is available.
- Lexicon size
  The dictionary contains 70,000 words.
- Domain coverage
  Both general terminology and computer industry terminology dictionaries are included.
- Morphology
  It appears that no morphology information is provided.
- Accuracy
  The dictionary is clearly manually constructed or verified.
- Translation preference information
  It is not clear whether translation preference information is present.
- Unicode compatibility
  There is no information available about the internal character code that is used.

Part IV. Machine Translation Resources

Definition. Modular resources for producing translations in both directions between English and Chinese. For languages that presently lack available bidirectional machine translation systems, unidirectional systems are also of interest. Both quick-and-dirty gloss-style translations and best-possible-quality translation systems are of interest.

Evaluation Methodology. Church & Hovy (1993) identified three types of measures for machine translation evaluation: text-based, cost-based, and system-based. Text-based measures are further subdivided into sentence-based measures, comprehensibility measures, and post-editing measures. Sentence-based measures, computed by hand-scoring each translated sentence for attributes such as semantic and stylistic correctness, report the fraction of the sentences assessed as satisfying each quality level. Comprehensibility measures are outcome-based document-level measures that use techniques such as multiple choice tests to determine whether representative end users were able to discern information contained in the original document by examining the translation. Post-editing measures are task-based measures that seek to characterize the difficulty of editing MT output to produce high-quality translations. Because that task is not appropriate to cross-language IR, post-editing measures are not considered further. Sentence-based measures and comprehensibility measures were both used in the DAPRA MT evaluations described by White & O'Connell (1994), and a strong correlation between the two types of measures was observed (Taylor, personal communication). White & Taylor (1998) report that the relationship between sentence-oriented measures and task performance has not yet been determined. Together, these results suggest that comprehensibility measures may be the better choice at present. Cost-based measures include measurements of the time required for translation and any marginal costs (such as human networked translation resources and human post-editing labor) associated with use of an MT system for the intended task. System based measures are glass box measures for the type described by Nirenburg, et al. (1996) that were explained in the section on "Other Term-Term Resources with Conceptual Structure." Oard & Resnik (1999) reported a relatively inexpensive technique for evaluating the use of gloss translations as indicative abstracts.

Evaluation Criteria.

- Availability
  - λ Available now
  - μ Projected to become available in 1999
- Cost
  - λ Free, or available with a multiuser license for $100 or less
  - ω Available for sale on negotiated terms or at a fixed price that exceeds $100 for multiple users
- Format
  - λ Available in an easily readable digital format
  - ω Available in a suitable format, but requires extensive preprocessing
  - μ Available only in hardcopy, data entry and validation estimated at over 40 hours
- Lexicon size
  - λ Large lexicon with good general news coverage available for translation in each direction
  - ω Moderate-sized lexicon with general news coverage available in at least one direction
- Translation speed
  - λ Nearly real time, at least 100 words per second on a high-end workstation
  - ω Suitable for online use, between 10 words per second and 100 words per second
  - μ Suitable for offline use, less than 10 words per second
- Browser compatibility
  - λ Works with the standard US version of Netscape or Internet Explorer
  - ω Requires special-purpose user interface software

Name	Availability	Cost	Format	Lexicon Size	Translation Speed	Browser Compatibility
Auto-Trans		ω	λ	λ		ω
Systrans	λ	ω	λ	ω	ω	ω
Transperfect	λ	ω	λ	ω		ω

Auto-Trans
A bidirectional English-Chinese machine translation system.

- Availability
  The software is described on a web site maintained by the ComStar Company, a retailer specializing in multilingual software, at http://www.gy.com/www/ww1/ww2/tong1.htm. It is not clear who makes the software, but the documentation suggests that it was developed in the Peoples Republic of China.
- Cost
  Although a description of the software is available, it appears to have been recently removed from the ComStar price list.
- Format
  The software runs under Windows 95, and the documentation implies that it is distributed on a set of diskettes.
- Lexicon size
  The lexicon claims to contain 2 million general terms and 30 million technical terms.
- Translation speed
  The translation speed is not specified.
- Browser compatibility
  Auto-Trans is configured for offline translation, rather than for use with a web browser.

Systran
A unidirectional Chinese to English machine translation system from Systran Software, Inc.

- Availability
  Presently available from Systran. Additional information is available at http://www.systransoft.com/pro.html.
- Cost
  A five-user license for Systran Professional Client/Server sells for $3,250 per bidirectional language pair, with prices increasing to $20,000 for a twenty-user license. A single-user standalone system is available for $1,000 per bidirectional language pair.
- Format
  Systran Professional is distributed on CDROM for Windows 95/NT.
- Lexicon size
  Systran Professional contains a large lexicon (2.5 million entries in 14 languages) that includes both general and technical terminology.
- Translation speed
  Systran Professional can translate about 40 words per second when used with a Pentium processor.
- Browser compatibility
  Browser support is not presently available for Chinese to English translation.

Transperfect
A unidirectional English to Chinese machine translation system from Otek International, Inc, a Taiwanese company.

- Availability
  Additional information is available in English at http://www.gy.com/www/ww1/ww2/amei1.htm and in Chinese at http://www.otek.com.tw/.
- Cost
  A single-user copy of Transperfect sells for $300 as a "professional edition" and $100 as a "standard edition." It is not clear how these versions differ.
- Format
  Transperfect runs under Windows 95, and is distributed as a set of diskettes.
- Lexicon size
  The lexicon contains about 100,000 words.
- Translation speed
  The translation speed is not specified.
- Browser compatibility
  Transperfect does not appear to be designed for use with web browsers.

References

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*. vol. 21, no. 4.

Comrie, Bernard. 1987. *The World's Major Languages*. (Croom Helm).

DARPA. 1995. *Proceedings Sixth Message Understanding Conference*. Columbia, MD. November.

DARPA. 1996. *Tipster Program Phase II*. Vienna, VA. May.

Davis, Mark. 1998. Unicode Normalization Forms. Draft Unicode Technical Report #15, The Unicode Consortium. August. Available at http://www.unicode.org/unicode/reports/tr15/.

Gallippi, Anthony F. 1996. Learning to Recognize Names Across Languages. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*. Copenhagen, Denmark. August.

Grishman, R. and B. Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*. Copenhagen, Denmark. August.

Hirschman, L. 1998. Language Understanding Evaluations: Lessons Learned from MUC and ATIS. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain. May.

Hovy, E. H. 1998. Creating Useful Evaluation Metrics for Machine Translation. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain. May.

Kikui, Gen-itiro. 1996. Identifying the Coding Scheme and Language of On-Line Documents on the Internet. In *Sixteenth International Conference on Computational Linguistics*. Copenhagen. August.

Knight, Kevin and Jonathan Graehl, 1997. Machine Transliteration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid. July. Available at http://www.isi.edu/natural-language/projects/GAZELLE.html.

McCullum, Sally and Monica Ertel. 1993. *Proceedings of the Second International Federation for Library Automation Satellite Meeting: Automated Systems for Access to Multilingual and Multiscript Library Materials*. Madrid, Spain. August.

Melamed, I Dan. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In *Third Workshop on Very Large Corpora*, Boston. Available at http://www.cis.upenn.edu/~melamed/.

Melamed, I. Dan. 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *2nd Conference on Emperical Methods in Natural Language Processing*, Providence, RI. Available at http://www.cis.upenn.edu/~melamed/

Oard, Douglas W. and Philip Resnik. 1999. Support for Interactive Document Selection in Cross-Language Information Retrieval. *Information Processing and Management*. To appear.

Padr∴, LluPs and LluPs MBrquez. 1998. On the Evaluation and Comparison of Taggers: The Effect of Noise in Testing Corpora. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal. August.

Reeder, F. and J. Geisler. 1998. Multi-Byte Issues in Encoding/Language Identification. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, Langhorne, PA. October.

White, J. S. and T. A. O'Connell. 1994. The DARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, MD.

White, J. S. and K. B. Taylor. 1998. A Task-Oriented Evaluation Metric for Machine Translation. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain. May.