WP1 Spec for 3rd TC-STAR SLT Evaluation

Summary

The third TC-STAR evaluation covers three translation directions:

- Mandarin to English (ZhEn)
- English to Spanish (EnSp)
- Spanish to English (SpEn)

Evaluation Campaign

Mandarin	• Text input: verbatim transcript • Speech input: single best + word graph
English and Spanish	• Text input: verbatim transcript + final text edition • Speech input: single best + word graph

Distinguish between different types of input for translation:

- final text edition (FTE)
  shown on webpage with multilingual translations
- verbatim transcriptions (Verbatim)
  with spontaneous speech phenomena: repairs, hesitations, word repetitions, grammatical errors
- recognizer output (ASR)
  in addition to the effects noted in the verbatim transcriptions, recognition errors both single-best and word graph output

Total number of test conditions:

Mandarin to English	1*Text : Verbatim	2 * speech single best + word graph
English to Spanish	2 * text : FTE + Verbatim	2 * speech : single best + word graph
Spanish to English	2 * text :FTE + Verbatim	2 * speech :single best + word graph

For each test condition:
25,000 running words for English (EPPS) and Mandarin (Voice of America)
50,000 running words for Spansih (25,000 from EPPS + 25,000 from Spanish Parliament)

Site commitments

For the evaluation for the various conditions in the table on the ELDA TC-STAR website (http://www.elda.org/en/proj/tcstar-wp4/tcs-slt-run3.htm)

DETAILS

Domains

- Mandarin to English:
  - ASR: broadcast news
  - SLT: NIST large data track (LDC data)
- Spanish to English
  - EPPS both for ASR and SLT
  - Spanish parliament both for ASR and SLT
- English to Spanish:
  - EPPS both for ASR and SLT

Conditions

ESEN	Primary – EPPES-Only Track	Secondary – Public Data Track
purpose	have strictly comparable systems	exploit special research topics push performance to the limits
input	- Final Text Edition - Verbatim transcription - Single best (ASR) + word graph	*as primary*
training	EPPS (bilingual, monolingual). This comprises the data from April 1996 to September 2004, December 2004 to May 2005 and December 2005 to May 2006. No additional bilingual data is allowed. Monolingual tools (e.g. POS-taggers) and publicly available monolingual data can be used.	*as primary, plus:* any publicly available data (but no private data) before the cut-off date (May 2006). Some additional corpora have already been made available on the RWTH TC-Star web page. These are the EU Bulletin Corpus, the JRC-Acquis Corpus and the UN Corpus. Participants are however not restricted to this additional data.
development	EPPS (bilingual) and Spanish Parliament (bilingual)	*as primary*
test	EPPS (bilingual) and Spanish Parliament (bilingual)	*as primary*

ENES	Primary – EPPES-Only Track	Secondary – Public Data Track
purpose	have strictly comparable systems	exploit special research topics push performance to the limits
input	- Final Text Edition - Verbatim transcription - Single best (ASR) + word graph	*as primary*
training	EPPS (bilingual, monolingual). This comprises the data from April 1996 to September 2004, December 2004 to May 2005 and December 2005 to May 2006. No additional bilingual data is allowed. Monolingual tools (e.g. POS-taggers) and publicly available monolingual data can be used.	*as primary, plus:* any publicly available data (but no private data) before the cut-off date (May 2006). Some additional corpora have already been made available on the RWTH TC-Star web page. These are the EU Bulletin Corpus, the JRC-Acquis Corpus and the UN Corpus. Participants are however not restricted to this additional data.
developmen t	EPPS (bilingual)	*as primary*
test	EPPS (bilingual)	*as primary*

ZHEN	Primary – Public Track	Secondary – Open Track
purpose	have strictly comparable systems	exploit special research topics push performance to the limits
input	- Verbatim transcription - Single best (ASR) + word graph	*as primary*
training	All public data available from LDC is allowed (except the corpus the test data is extracted from, LDC2002T01). Monolingual tools are allowed.	All data is available (except the corpus the test data is extracted from, LDC2002T01)
developmen t	VoA data (bilingual)	*as primary*
test	VoA data (bilingual)	*as primary*

Material distributed for the evaluation campaign

Language Resources for Mandarin to English

Two input types:

- text input (verbatim)
- speech input: single best + word graph

Training Data

Training data is the same as training data of the run #1:

LDC Code	Name
LDC2003E14	FBIS Multilanguage Texts
LDC2004E12	UN Chinese English Parallel Text Version 2
LDC2004T08	Hong Kong Parallel Text
LDC2002E17	English Translation of Chinese Treebank
LDC2002E18	Xinhua Chinese-English Parallel News Text Version 1.0 beta 2
LDC2002L27	Chinese English Translation Lexicon version 3.0
LDC2003E01	Chinese-English Name Entity Lists version 1.0 beta
LDC2005E47	Chinese English News Magazine Parallel Text
LDC2002T01	Multiple-Translation Chinese (MTC) Corpus
LDC2003T17	Multiple Translation Chinese (MTC) Part 2
LDC2004T07	Multiple Translation Chinese (MTC) Part 3
LDC2005T06	Chinese News Translation Text Part 1
LDC2005T01	Chinese Treebank 5.0
LDC2003E07	Chinese Treebank English Parallel Corpus

Development and Evaluation Data

- Dev : previous development/evaluation data (dev05, eval05, dev06, eval06)
- Test data: 25 000 words from TDT3, Dec 1998
- Segments of man transcripts will be translated by agencies
  - According to ELDA translation guidelines
  - Two reference translations per segment
  - Ref translations data format of NIST/LDC (see deliverable D4)
- Auto transcripts will be provided by LIMSI and UKA
  - Single outputs + rover of two systems
  - Output at character level with time stamps
  - Data format of NIST/LDC (see deliverable D4)
  - Translations of auto transcripts will be aligned automatically

Language Resources for English to Spanish

Training Corpora

Provided resources are:

EPPS English verbatim transcriptions May 2004- Jan 2005	Transcribed by RWTH
EPPS English final text edition April 1996 to Sept 2004	Provided to TCSTAR by RWTH
EPPS English final text edition Dec 2004 - May 2005	English and Spanish parallel texts are aligned.
EPPS English final text edition Dec 2005 - May 2006	English and Spanish parallel texts are aligned.
EU Bulletin Corpus	The Bulletin of the European Union provides an insight into the activities of the European Commission and the other Community institutions." It is published in a monthly basis and parallel versions in Spanish and English are available up to 2004. The corpus is available in a raw version, with the html documents as downloaded from the pages of the European Union, or as a sentence aligned version. Sentence alignment provided by RWTH
JRC-Acquis Multilingual Parallel Corpus	Before joining the European Union (EU), the new Member States (NMS) needed to translate and approve the existing EU legislation, consisting of selected texts written between the 1950s and 2005. This body of legislative text, which consists of approximately eight thousand documents and which covers a variety of domains, is called the Acquis Communautaire (AC)." The RWTH conducted an additional sentence level alignment of the corpus. Provided by RWTH
UN Parallel Corpus	The text files published in this corpus were provided to the LDC by the United Nations in New York, for use by the research community in developing machine translation technology. This material has been drawn from the UN's electronic text archives covering the period between 1988 and (portions of) 1993." We are not allowed to distribute this data, therefore a set of tools has been made available for carrying out the sentence alignment by each partner. Alignment tools provided by RWTH

Development and Evaluation Data

A subset of EPPS data, defined by ELDA in cooperation with WP2.

- Development: development and data from first and second years (dev05, eval05, dev06, eval06)
- Evaluation: a subset after the May 2006 cut-off date.

For the secondary condition, the first date is the cut-off date for any training material.

There are only original speakers (i.e. politicians) in the material.

Two reference translations for the final text edition and two for the verbatim transcription (which will be also used as reference for ASR evaluation).

Spanish and English test material are not translations of each other.
Amount: about 25 000 running words or about 3 hrs of speech.

Language Resources for Spanish to English

Training Corpora

Provided resources are:

EPPS Spanish verbatim transcriptions May - Jan 2005	Transcribed by UPC
EPPS Spanish final text edition April 1996 to Sept 2004	Provided to TCSTAR by RWTH
EPPS Spanish final text edition Dec 2004 - May 2005	English and Spanish parallel texts are aligned.
EPPS Spanish final text edition Dec 2005 - May 2006	English and Spanish parallel texts are aligned.
EU Bulletin Corpus	The Bulletin of the European Union provides an insight into the activities of the European Commission and the other Community institutions." It is published in a monthly basis and parallel versions in Spanish and English are available up to 2004. The corpus is available in a raw version, with the html documents as downloaded from the pages of the European Union, or as a sentence aligned version. Sentence alignment provided by RWTH
JRC-Acquis Multilingual Parallel Corpus	Before joining the European Union (EU), the new Member States (NMS) needed to translate and approve the existing EU legislation, consisting of selected texts written between the 1950s and 2005. This body of legislative text, which consists of approximately eight thousand documents and which covers a variety of domains, is called the Acquis

	Communautaire (AC)." The RWTH conducted an additional sentence level alignment of the corpus. Provided by RWTH
UN Parallel Corpus	The text files published in this corpus were provided to the LDC by the United Nations in New York, for use by the research community in developing machine translation technology. This material has been drawn from the UN's electronic text archives covering the period between 1988 and (portions of) 1993." We are not allowed to distribute this data, therefore a set of tools has been made available for carrying out the sentence alignment by each partner. Alignment tools provided by RWTH

Development and Evaluation Data

A subset of Spanish Parliament data (cortes), defined by ELDA in cooperation with WP2.

- Development: development and data from first and second years (dev05, eval05, dev06, eval06)
- Evaluation: a subset after the May 2006 cut-off date.

A subset of EPPS data, defined by ELDA in cooperation with WP2.

- Development: development and data from first and second years
- Evaluation: a subset after the May 2006 cut-off date.

For the secondary condition, the first date is the cut-off date for any training material.

There are only original speakers (i.e. politicians) in the material.

Two reference translations for the final text edition and two for the verbatim transcription (which will be also used as reference for ASR evaluation).

Spanish and English test material are not translations of each other.
Amount: about 50 000 running words or about 6 hrs of speech.

Formats and Tools

Character Encoding

Chinese to English: GB2312 for Chinese, UTF-8 for English
English to Spanish and Spanish to English: UTF-8 throughout all text EPPS and CORTES related data.

Topics Like Enriched Transcriptions and Punctuation Marks

- punctuation marks: use of punctuation marks for Final Text Edition and Verbatim transcriptions and ASR
- true case
- orthographic conventions as used for speech recognition
- format: the same format as used for the NIST evaluations (both for Chinese and for EnEs)
- entities like e.g. dates are currently represented in different ways for the Final Text Edition and the Verbatim transcription, e.g.
  FTE: January 27, 1999 (preferred format)
  VT: January twenty-seventh nineteen ninety nine
  The assumption is that each group can process these different formats appropriately

Scoring Software

- on the webpage: http://www.elda.org/en/proj/tcstar-wp4/tcs-slt-run3.htm
- Sample files to check proper installation (on webpage).

Word Graphs

- word graphs produced by the ASR partners must be made available to the consortium
- there was no consensus about the use of word-graphs in this evaluation
- if they are to be used, word graphs for development data are to be exchanged on a bi-lateral basis among the groups
- there will be no common word graph format
- besides word graphs, complete word lists required

Evaluation Procedure

Evaluation Measures

- automatic scores and general subjective evaluation
- BLEU/NIST, BLEU/IBM, NIST, mWER, mPER, WNM
- A common software package is available on the ELDA webpage

Subjective Evaluation

- important to assess quality of systems (fluency>3? Adequacy>3?)
- will be performed on a subset of carefully selected runs
- for English to Spanish direction only
- specifications: to come (preferably the same than 2nd year)

Sequential Organization and Evaluation Schedule

First ASR evaluation, which also produces the single best and word graph results (for En, Es); followed by SLT evaluation.

Schedule for SLT:

Sept. 06, 2006 ELDA sends Dev data:
Jan. 31, 2007 ELDA sends Test data:
Feb. 07, 2007 Deadline for sending back translations to ELDA
Feb. 09, 2007 Automatic preliminary results
Feb. 23, 2007 Automatic final results
Mar. 16, 2007 Human evaluation results