Technology and Corpora for Speech to Speech Translation
http://www.tc-star.org

Project no.: **FP6-506738**

Project Acronym: **TC-STAR**

Project Title: **Technology and Corpora for Speech to Speech Translation**

Instrument: **Integrated Project**

Thematic Priority: **IST**

Deliverable no.: D4
Title: SLT Baseline & Specifications

Due date of the deliverable:	30^{th} of September 2004
Actual submission date:	22^{nd} of November 2004
Start date of the project:	1^{st} of April 2004
Duration:	36 months
Lead contractor for this deliverable:	ITC-irst
Authors:	M. Federico (ITC-irst), M. Koss (UKA), S. Roukos (IBM), R. Zens (RWTH)

Revision: [version 3.1]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium(including the Commission Services)	
CO	Confidential, only for members of the consortium(including the Commission Services)	

D4 SLT: Baselines & Specifications

Contents

1 Introduction

According to the TA, baseline specifications and systems for Spoken Language Translation (SLT) should be ready by month 6, while the first evaluation will take place on month 11.
At the Kick-off meeting on May 13, 2004, it was decided that baseline systems of SLT WP would be developed and evaluated on Mandarin-to-English only. It was also proposed that in the first formal evaluation (month 11), SLT systems for Mandarin-to-English, Spanish-to-English and English-to-Spanish will be evaluated. As a consequence, English-Spanish SLT systems developed for the first evaluation will be taken as reference baselines for successive evaluations.

This report defines specifications and evaluation conditions of Mandarin-to-English baselines systems developed and evaluated at month 6 [1]. Moreover, results and descriptions of each baseline are reported, too.

Baselines systems were supposed to meet specifications/requirements on the following issues: (i) input/output format, (ii) evaluation conditions.

To assess progress over time, it is important that baseline systems can be run over new test data as well, so that the most recent SLT system can be fairly compared against the baseline, over different test sets. This implies that each partner properly maintains his baseline system over all the duration of the project.

Clearly, we also expect that future SLT systems will be able to operate on different (richer) input formats, e.g. word lattices, n-best lists, punctuation rich format, case sensitive format, etc.. Hence, specifications should chose an input format for SLT that future ASR systems will always be able to produce.

Concerning the output format, a similar story applies, because SLT will provide input to the speech synthesis module. In addition, output format specifications must be set so that consistent automatic and subjective evaluation of baselines and systems can be performed over time. Hence, the output format should be chosen in a way so that it can be easily reproduced by future systems and accepted by future automatic scoring programs.

This report is organized as follows. First, we survey language resources available for developing and evaluating Mandarin-to-English SLT systems. Then, specifications for baseline system evaluation are defined, namely, input/output format and training conditions. A section follows which describes the applied evaluation methods, evaluation conditions, and the schedule of the evaluation. In the same section, results of the baseline system evaluation are presented and discussed. Finally, a section is devoted to the description of the baseline systems that took part in the evaluation.

2 Survey of Language Resources

At this time the largest repository of Mandarin and English resources is maintained by the Linguistic Data Consortium (LDC).

2.1 Parallel resources

An excellent reference for training data configuration is given by the NIST MT evaluation campaigns. In particular, the so called "Large Data Condition" track defines a list of publicly available resources that can be used to develop text MT systems from Mandarin to English.

[1] In principle, apart from using different data sets, the same conditions and specifications will apply to the English-Spanish baselines.

Unfortunately, there do not seem to exist large parallel corpora of transcribed Chinese-Mandarin broadcast news.

2.2 Monolingual resources

The so called English Gigaword (LDC2003T05) and Chinese Gigaword (LDC2003T09) are available. Monolingual corpora of English and Mandarin broadcast news transcripts (both manual and automatic) can be found within "TDT3 Multilanguage Text Version 2.0". Among others, this corpus contains 3,800 stories transcribed from "VOA[2] Mandarin Chinese news programs".

2.3 Multiple translations

In order to set-up a benchmark test for SLT baselines, there is need for a collection of automatically and manually transcribed Mandarin broadcast news, ideally provided with multiple translations.
A sample of such data is "indirectly" available from LDC, namely from the following two databases:
– "Multiple-Translation Chinese Corpus" (LDC2002T01)
– "TDT3 Multilanguage Text Version 2.0" (LDC2001T58)

The first corpus, among others, contains 26 stories (for a total of 9,256 Chinese characters) extracted from "Voice of America Mandarin" broadcast transcripts. For each story, 11 translations produced by humans are available. The second corpus, contains automatic and manual transcripts of the same stories. Automatic transcripts were produced by the Dragon Mandarin speech recognizer.
Remark: the VOA Mandarin transcripts in TDT3 were created manually by a professional transcription service, but with limited editorial quality control – while generally quite complete, these transcripts were not expected to exceed the quality or accuracy of closed-caption text in television broadcasts.

3 Baseline Specifications

3.1 Input/Output format

SLT baselines should be able to work with the simplest input that can be produced by every ASR system:

- single-best hypothesis,
- case-insensitive,
- no punctuation information

Moreover, standard formats of numbers and abbreviations should be defined beforehand.
Text translation is usually applied to syntactically well formed text segments, which are then evaluated individually with respect to a set of target references. A similar principle could be applied on spoken documents, either by assuming a manual syntax-driven segmentation, or an automatic segmentation based on acoustic information.
Similarly, SLT baselines must be able to produce text output without case nor punctuation information.

[2] Voice of America

The rationale for this is that even if progress in ASR will be made in terms of producing richer input formats for SLT, the simpler input will always be available, so that it can be passed to the baseline system.
Further, if progress in SLT systems will be made in terms of producing richer and more readable output, new SLT systems are also expected to improve translation quality at the basic level, i.e. to compute more fluent and more adequate word strings. Of course, new SLT systems can use whatever input format suites best for them.
A potential side-effect to be investigated is the possible lower correlation between human and automatic scores that could occur if case information and punctuation are not considered. This issue is relevant to SLT, given that such linguistic information is missing in the source signal. There are two possible alternative solutions to this problem: either the speech recognizer should provide an enriched transcription or the SLT should be able to produce translations with capitalization and punctuation even if the input does not contain such information.
For the sake of completeness, the first evaluation of SLT baseline systems included a track assuming enriched transcriptions in input. Finally, layout of input/output of SLT baselines followed the conventions adopted in the 2004 NIST MT evaluation (see Appendix).

3.2 Training conditions

Baselines should be trained on a specified set of publicly available resources. This is an important requisite in order to compare performance of partners' systems under fair and controlled conditions. Moreover, this will give experiments of partners a higher relevance and also improve assessment of methods. Future evaluations will possibly enlarge or modify the list of training resources defined for the baseline systems.
The list of resources used to train the Mandarin-Chinese baselines is reported in Appendix.

4 Baseline Evaluation

SLT baseline systems have been evaluated on a benchmark of 26 broadcast news stories from VOA, for a total of 9,256 Chinese characters. For each story, automatic and manual Mandarin transcriptions are available, as well as 11 human-made translations.
Automatic evaluation was performed in a centralized way, in order to ensure consistency of results. Participants, after receiving human and automatic transcripts to be translated, had to send back results to the site managing the evaluation, namely ITC-irst.

4.1 Evaluation method

Runs were evaluated with the following automatic scores:

- BLEU: the geometric mean of n-gram precision by the system output with respect to the reference translations.
- NIST: a variant of BLEU using the arithmetic mean of weighted n-gram precision values.
- mWER: multiple word error rate, i.e. the edit distance between the system output and the closest reference translation.
- mPER: position independent word error rate, i.e. a variant of MWER which disregards word ordering.

According to the kind of evaluation condition, namely plain versus enriched transcription, the following settings were used in the scores:

- case insensitive and no punctuation
- case sensitive and punctuation

Unfortunately, the alignment between automatic transcripts and manual transcripts of the benchmark was originally only at the document level. In other words, the sentence-wise segmentation in the manual transcripts has no equivalent in the automatic transcripts. This of course impacts on the evaluation scores, which rely on multiple reference translations at the sentence level. To solve this problem two alternative options were investigated before releasing the test data.

Option 1. Sentence boundaries in the human transcripts are manually mapped into the automatic transcripts. This operation must be performed by a Chinese native person.

Option 2. New sentence boundaries are defined on the automatic transcripts, e.g. based on the presence of pauses. Then, automatic evaluation is performed at the document level rather than at the sentence level. This result in different automatic scores, which however should be well correlated with the sentence-level evaluation. To verify the effectiveness of Option 2, we performed automatic evaluation measurements on 20 system runs which participated in the 2003 NIST MT evaluation campaign. Correlations of scores between the two modalities of evaluation (sentence-level vs document-level) resulted as follows:

Score	*Corr. Coeff.*
BLEU	0.998194
NIST	0.997088
MWER	0.982258
MPER	0.968336

Finally, thanks to the availability of a Chinese native person at RWTH, Aachen, we opted for Option 1. In particular, the ASR texts were manually annotated with sentence boundaries and punctuation. The presence of punctuation in the source should simulate an ASR system able to produce enriched transcriptions.

4.2 Evaluation conditions and schedule

Two main experimental conditions have been defined, each of which includes a primary and a contrastive condition.

A.	Plain Transcription	
	A.1.	Primary Condition
		input: ASR transcription without punctuation
		output: no punctuation, no case information
	A.2	Contrastive Condition
		input: human transcription without punctuation
		output: no punctuation, no case information
B.	Rich Transcription	
	B.1	Primary Condition
		input: ASR transcription with punctuation
		output: punctuation and case information
	B.2	Contrastive condition
		input: human transcription with punctuation
		output: punctuation and case information

In addition to the above, partners were allowed to submit runs under any other condition they liked. Given that all partners interested in Mandarin-to-English translation already own baseline systems more or less developed according to the here considered requirements, the following schedule was finalized:

Sep 28, 2004	- Availability of benchmark test by ITC-irst
Oct 11, 2004	- Submission of runs by all partners to ITC-irst
Oct 11, 2004	- Partners send description of baseline to ITC-irst
Oct 15, 2004	- Release of Report on SLT baselines by ITC-irst

To comply with copyright issues, before receiving the benchmark from ITC-irst, partners were asked to confirm by e-mail to that they already had received from LDC the two databases:

- LDC2002T01 - "Multiple-Translation Chinese Corpus
- LDC2001T58 - "TDT3 Multilanguage Text Version 2.0"

4.3 Evaluation results

Runs for each evaluation condition were prepared and submitted by the following partners: RWTH, UKA, IBM, and ITC-irst. Performance evaluation by means of automatic scores was centrally performed at ITC-irst.
Before computing the automatic scores in the plain transcription conditions (A1 and A2), both human reference translations and the target transcriptions and system outputs were filtered in order to remove punctuation marks. In particular, the very same text tokenization performed by the BLEU script was applied, and finally isolate punctuation marks were removed.
Automatic scores were computed by a tool kindly provided by the USC/Information Sciences Institute. For each score, the tool also provides .95% confidence intervals, which are computed with a bootstrap technique. Results for all runs are shown in Table 1 and Table 2. Statistically significant differences between systems are indicated with small indexes close to each score. In

particular, indexes indicate which systems performed significantly worse than a given system, i.e. the corresponding confidence intervals do not overlap.

	A1 Primary				A2 Contrastive			
	BLEU	NIST	WER	PER	BLEU	NIST	WER	PER
IBM_1	22.26	7.930_2	87.70	59.16	25.40	8.708	82.19	52.90
$IRST_2$	21.12	7.178	90.32	63.01	25.47	8.283	84.03	55.40
$RWTH_3$	$30.21_{1,2}$	8.499_2	79.45	53.48_2	$35.26_{1,2}$	9.386_*	73.61	47.66_2
UKA_4	$28.34_{1,2}$	7.984_2	76.57_2	54.06_2	$32.02_{1,2}$	8.669	73.13_2	49.58

Table 1: Results under the **plain transcription** condition with automatic transcriptions (A1 primary) and manual transcriptions (A2 contrastive). Indexes indicate systems which performed worse (α=0.05) according to the same score. $*$ is a shorthand for all remaining systems.

	B1 Primary				B2 Contrastive			
	BLEU	NIST	WER	PER	BLEU	NIST	WER	PER
IBM_1	20.79	7.736	85.59	59.22	23.66	8.503	80.28	53.57
$IRST_2$	19.89	7.072	86.62	62.59	23.77	8.139	80.87	55.47
$RWTH_3$	$33.90_{1,2}$	$8.573_{1,2}$	$73.12_{1,2}$	$50.57_{1,2}$	39.56_*	9.429_*	$67.93_{1,2}$	$45.53_{1,2}$
UKA_4	$30.65_{1,2}$	$8.146_{1,2}$	$74.65_{1,2}$	53.52_2	$33.59_{1,2}$	8.819_2	71.78	49.36

Table 2: Results under the **rich transcription** condition with automatic transcriptions (B1 primary) and manual transcriptions (B2 contrastive). Indexes indicate systems which performed worse (α=0.05) according to the same score. $*$ is a shorthand for all remaining systems.

4.4 Analysis and conclusions

An analysis of the results shows that in general the systems of UKA and RWTH perform better than those of IBM and ITC-irst. Moreover, as many large differences in scores are not statistically significant, we might easily conclude that the used sample is too small for the purpose of ranking system performance.

In fact, experience with this evaluation rather suggests two kinds of possible improvements: (i) using a larger test set and (ii) applying sharper significance tests. For the second issue, we believe that significance tests should directly address score differences of system pairs, and also take advantage of the fact that systems are run on the same test data, similarly to how paired-sample mean tests do.

For what concerns the test set size, a preliminary analysis was made by joining runs submitted under condition B2, in a way to simulate larger test sets. The aims was to check the reductions in size of the 95% confidence intervals of the BLEU score. The four B2 systems, evaluated independently, give on-average BLEU confidence-intervals of size 3.7. After joining the outcomes of all systems as it would be one single run, the BLEU confidence interval size reduces to 2.2, which is almost a 40% relative reduction. In other words, a test set four times larger could probably provide 40% more accurate BLEU scores. In other words, a BLEU score difference between two systems of a about two points would result statistically significant, according to the here adopted significance test.

Other preliminary work investigated the need for many alternative reference translations. Again, all systems of the B2 condition were evaluated with the BLEU score by varying the number of reference translations. Results are reported in Table 3 in terms of 95%-confidence intervals of

	1 references	3 references	5 references	7 references	9 references	11 references
IBM	8.318-10.48	14.18 - 16.89	17.46 - 20.37	19.20 - 22.18	20.83 - 23.83	22.15 - 25.28
IRST	8.111-10.39	14.60 - 17.44	16.82 - 20.00	19.01 - 22.30	20.46 - 23.83	21.94 - 25.55
RWTH	13.66-16.69	23.77 - 27.48	28.95 - 32.99	33.14 - 37.27	35.4 - 39.53	37.36 - 41.63
UKA	11.52 - 14.3	20.72 - 24.13	24.6 - 28.3	27.6 - 31.58	29.37 - 33.58	31.41 - 35.76

Table 3: Confidence intervals of BLEU scores in the B2 condition by varying the number of reference translations.

BLEU scores. It basically results that BLEU scores significantly increase with the availability of more reference translations, but at the same time the ranking of systems basically remains unaltered. Moreover, the sizes of the confidence intervals do not seem to be affected by the number of reference translations.

As a final consideration, in order to get more precise scores, future evaluations should make use of sharper significance tests and more test data, at least four times the amount used for evaluating baseline systems. As the reference translations of the current test data contain on average 6,000 running words, without punctuation, around 25,000 words should be used in future evaluations. The increase of data can be well compensated by reducing the number of reference translations to, e.g., 3 for each sentence.

5 Baseline Descriptions

5.1 IBM: System Description

IBM Chinese-to-English baseline system is a phrase-based statistical machine translation system, comprising of 2.1 million phrase pairs. Phrases are extracted from word alignments, and used for translating input sentences along with a trigram language model. The phrase translation model and language model are trained on lower-cased English. Chinese input is word-segmented and tagged for entities such as numbers and dates before decoding. Casing is recovered from the decoding output by a language-model based true-caser.
The following subsections will discuss the word alignment procedure, the phrase extraction algorithm, and the DP-based decoding.

5.1.1 Translation Models and Word Alignments

Translation Model Traditional IBM Models, for example IBM Model 1, start with an all-inclusive translation model by allowing all co-occurring translations. The training then shrinks the model by making the probabilities of bad translations smaller and smaller. In the end, a near-zero effectively eliminates the translation. Although in the end, some really bad translations do go away, the models, in general, are quite noisy. In contrast, we start our training by building a small base model (as explained below). We then word align the parallel corpus using this base model. Only the aligned word pairs enter into the final translation model. Once a pair is aligned, it is taken out of the sentence pair. The remaining of the sentence pair, the reduced sentence pair as we call them, become the training data for the next iteration. We call this training incremental training. In each iteration, more aligned word pairs enter into the final model. In this way, the training grows the model.
The base model is built by matching words according to the formula below, similar to the i-Divergence formula:

$$\frac{N_e N_f}{N} - N_{ef} + N_{ef} \cdot \log \frac{N_{ef} N}{N_e N_f}$$

N_e is the number of sentences in which the English word e occurs, N_f is the number of sentences in which the foreign word f occurs, N_{ef} is where they both occur; and N is the total number of sentences.
For each foreign word f, we compute the above score for all co-occurring English word e. We then put the K highest ranked English words into our base model.

Word Alignment Now that we have a translation model we move on to the task of aligning the parallel corpus. In the same way as we build the translation model, our main objective in word alignments is high precision. The algorithmic aspect of the alignment model is given below.

```
Input:  English sentence E {e1, e2, ..., en}
        Foreign sentence F {f1, f2, ..., fm}
          foreach f in F {
                e* = f's best translation among E
                f* = e*'s best translation among F
                if (f* == f)
                     align(f, e*)
          }
```

The idea is that only bi-directionally best word choices are aligned.

5.1.2 Phrase Extraction

We view the word alignments as a Boolean matrix. Our phrase extraction is a projection-based procedure on this matrix. An example of the alignment matrix is shown below where an x marks where the words are aligned.

	e_1	e_2
f_1	x	
f_2		x

Figure 1

	e_1	e_2	e_3
f_1	x		
f_2			x

Figure 2

For a given range of foreign words $[f_i \dots f_j]$ $(i \leq j)$, we find the English word range according to the word alignments $[e_k \dots e_l]$ $(k \leq l)$. For example, given $[f_1, f_2]$, alignment 1 (in Figure 1) gives English range $[e_1, e_2]$, alignment 2 (Figure 2) gives $[e_1, e_3]$. That is, the foreign word range is projected to the English word range. We then check the English word range to see if it has any "holes" in it. In the first example, there isn't any because all words in $[f_1, f_2]$ are connected to all words in $[e_1, e_2]$. The following phrase pair is then extracted:

$$f_1 f_2 \longrightarrow e_1 e_2$$

In the second case, however, there is a hole on the English side, namely that e_2 is unconnected to either f_1 or f_2. We do not always want to extract phrases with holes in them especially when e_2 is a content word. If e_2 is a function word (like prepositions, determiners, etc.) we will extract this phrase pair:

$$f_1 f_2 \longrightarrow e_1 e_2 e_3$$

the intuition is that function words behave like glues and we will allow it to be included. Content words, however, are required to be aligned in order for them to be part of a phrase translation. It is possible to have the situation in Figure 3. When projecting $[f_1, f_3]$ we get $[e_1, e_2]$ because f_3 is not aligned. Just as we are careful about "holes" on the English side, we want to do the same on the foreign word side. To achieve this, once we have the projected English word range, we project that back to the foreign side. Projecting $[e_1, e_2]$ back gets us $[f_1, f_2]$ which is not the original input range $[f_1, f_3]$. In this case we again

	e_1	e_2
f_1	x	
f_2		x
f_3		

Figure 3

check to see if the "hole" (f_3) is function word or not. If it is we will allow it and phrase pair will be extracted

$$f_1 f_2 f_3 \longrightarrow e_1 e_2$$

We apply this procedure to every foreign word range and every sentence pair in the training data.

5.1.3 Decoding

Phrase unigram model and word trigram language model, are used for decoding. Phrase unigram probability is defined below, where n is the number of distinct phrases and b denotes a phrase:

$$p(b) = \frac{count(b)}{\sum_{i=0}^{k} count(b_i)}$$

Word trigram language model probability is computed at target phrase boundaries only, skipping over words within a target phrase in case the target phrase length is longer than 2 words. Trigram language model probability between adjacent target phrases is computed, as shown below.

$$p(\bar{e}_i \mid \bar{e}_{i-1}) = p(e_l \mid e_h, e_{h-1})$$

$\bar{e}_i$ is the current target phrase, $\bar{e}_{i-1}$is the previous (one or more) target phrase in the hypothesis. e_l is the first word of $\bar{e}_i$, e_h the last target word in the hypothesis and e_{h-1} the second to the last target word in the hypothesis. The task of the decoder is to find the phrase sequence that maximizes the product of the unigram phrase probability and the trigram language model probability.
In decoder implementation, we use a DP-based beam search procedure, as discussed in [Tillmann 2003]. We start with an initial empty hypothesis. We maximize over all phrase segmentations $b_{1,n}$, where n is the number of phrases covering the input sentence, with the source phrases yielding a segmentation of the input sentence, generating the target sentence simultaneously. The decoder processes the input sentence 'cardinality synchronously', i.e. all partial hypotheses active at a given point cover the same number of input words. We prune out weaker hypotheses based on the cost (for phrase unigram probability and trigram language model probability) they incurred so far. The cheapest final hypothesis - the hypothesis with the highest probability - with no un-translated source words is the translation output.

References

[Tillmann 2003] C. Tillmann. A projection extension algorithm for statistical machine translation. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. pages 1-8, Philadelphia, 2003

5.2 ITC-irst: System Description

The ITC-irst SMT system implements a log-linear model which combines feature functions resulting from an extension of the IBM Model 4 [5] to *phrases*. The use of phrases rather than words has recently emerged as a mean to cope with the limited context that IBM models exploit to guess word translation (lexicon model) and word positions (distortion model) [1, 4, 2, 3].

5.2.1 Log-linear Model

Given a source string $\mathbf{f}$ and a target string $\mathbf{e}$, the framework of maximum entropy [6] provides a mean to directly address the posterior probability $\Pr(\mathbf{e} \mid \mathbf{f})$. By introducing the hidden *alignment* variable $\mathbf{a}$, the usual SMT optimization criterion is expressed by:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \approx \arg\max_{\mathbf{e},\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \tag{1}$$

The conditional distribution $\Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f})$ is determined through suitable real valued features functions $h_i(\mathbf{e}, \mathbf{f}, \mathbf{a}), i = 1 \ldots M$, and takes the parametric form:

$$p_\lambda(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) = \frac{\exp\{\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}, \mathbf{a})\}}{\sum_{\mathbf{e},a} \exp\{\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}, \mathbf{a})\}} \tag{2}$$

Feature functions are computed by means of log-models [7] estimated on a sample of phrase-pairs. Phrase-pairs are extracted from a so called *union alignment* between sentence pairs of the training corpus [3]. The weights of the interpolation are optimized through a training procedure which directly aims at minimizing translation errors on a development set [8].

Feature functions are logarithms of 6 probability models, which define an extension of IBM Model 4 to phrases:

- 1 language model at the word level
- 2 fertility models at the phrase level
- 2 distorion models at the phrase level
- 1 lexicon model at the phrase level

5.2.2 Decoding Algorithm

Given the source sentence $\mathbf{f}$, the optimal translation $\mathbf{e}^*$ in equation (1) is computed by dynamic programming through a recursive formula which expands previously computed partial theories, and recombines the new expanded theories. A theory can be described by its *state*, which only includes information needed for its expansion; two partial theories sharing the same state are identical (undistinguishable) for the sake of expansion, i.e. they should be recombined.
To limit the number of theories to generate, pruning methods and search constraints are introduced, which mildly impact on the optimality of the search algorithm.

- *threshold pruning*: partial theories whose score is smaller than the current optimum score times a given factor are eliminated.
- *histogram pruning*: hypotheses not among the top N best scoring ones are pruned.
- *reordering constraint*: each expanded theory must cover one one of the first 4 empty positions in the source string, from left to right.

5.2.3 Chinese Segmentation

Given a sequence of Chinese characters x_1^n, word segmentation is the task of guessing the number of words c contained and the corresponding word boundary positions $n_1^c = n_1,\ n_2 \ldots n_c$. From a statistical perspective, we look for the segmentation maximizing the text log-likelihood:

$$L^*(x_1^n) = \max_{c,n_1^c} L(x_1^n; c; n_1^c) = \max_{c,n_1^c} \sum_{i=1}^{c+1} \log P(w = x_{n_{i-1}}^{n_i - 1}) \tag{3}$$

where $1 = n_0 < n_1 < n_2 < \ldots < n_c < n_{c+1} = n + 1$.
The maximization in eq. (3) is solved by dynamic programming. The probability $P(w)$ is computed by a word model which combines statistics gathered from a large segmented corpus. Namely, frequencies at level of words, word-lengths, and character n-grams.

5.2.4 Pre-/Post-processing

Preprocessing and postprocessing consist of a sequence of actions aiming at normalizing text and are applied both for preparing training data and for managing text to translate. The same steps can be applied to both source and target sentences, accordingly with the language. Input strings are tokenized, and put in lowercase. Text is labeled with few classes including cardinal and ordinal numbers, week-day and month names, years and percentages. Translation is performed in a case-insensitive modality. Case information is added subsequently by means of a statistical maximum entropy tagger [6]. The tagger is trained on a large monolingual corpus in the target language.

5.2.5 Training Data

Chinese word segmentation algorithm was trained on the `Mandarin.fre` word-frequency list distributed by LDC. Training of the translation made use of the following parallel corpora:

LDC2002E17	- English Translation of Chinese Treebank
LDC2004E09	- Hong Kong Hansard Parallel Text
LDC2003E25	- Hong Kong News Parallel Text (1997-2003)
LDC2002L27	- Chinese English Translation Lexicon
LDC2002E58	- Sinorama Chinese-English Parallel Text
LDC2002E18	- Xinhua Chinese-English Parallel News Text

Monolingual resources were limited to the English part of the above parallel corpora. In the following there are statistics about the training data.

System	#sent.	# Man. Words	# Eng. words	Man dict	Eng Dict
Cond A	1.7M	32.6M	35.6M	51K	40K
Cond B	1.7M	36.5M	38.8M	51K	40K

References

[1] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China, October 2000, pp. 440–447.

[2] S. Vogel, Y. Zhang, F. Huang, A. Venugopal, B. Zhao, A. T. a nd M. Eck, and A. Waibel, "The CMU statistical machine translation system," in *Proc. of the Machine Translation Summit IX*, New Orleans, LA, 2003.

[3] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of HLT-NAACl 2003*, Edmonton, Canada, 2003, pp. 127–133.

[4] D. Marcu, "Towards a unified approach to memory- and statistical-based machine translation," in *Proc. of the 39th Meeting of the Association for Computational Linguistics (ACL)*, Toulouse, France, 2001, pp. 378–385.

[5] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–313, 1993.

[6] A. Berger, S. Della Pietra, and V. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[7] F. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *ACL02: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, PA, Philadelphia, 2002, pp. 295–302.

[8] M. Cettolo and M. Federico, "Minimum Error Training of Log-Linear Translation Models," in *Proc. of International Workshop on Spoken Language Translation (IWSLT)*, Kyoto, Japan, 2004, pp. 103-106.

5.3 RWTH: System Description

In statistical machine translation, we are given a source language ('French') sentence $f_1^J = f_1 \ldots f_j \ldots f_J$, which is to be translated into a target language ('English') sentence $e_1^I = e_1 \ldots e_i \ldots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\begin{aligned} \hat{e}_1^I &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I | f_1^J) \right\} \\ &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} \end{aligned}$$

This decomposition into two knowledge sources is known as the source-channel approach to statistical machine translation [1]. It allows an independent modeling of target language model $Pr(e_1^I)$ and translation model $Pr(f_1^J | e_1^I)$. The target language model describes the well-formedness of the target language sentence. The translation model links the source language sentence to the target language sentence. It can be further decomposed into alignment and lexicon model. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. We have to maximize over all possible target language sentences.

An alternative to the classical source-channel approach is the direct modeling of the posterior probability $Pr(e_1^I | f_1^J)$. Using a log-linear model [2], we obtain:

$$Pr(e_1^I | f_1^J) = \frac{\exp \left[\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right]}{\sum_{e'^I_1} \exp \left[\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J) \right]}$$

The h_m denote the feature functions. As a decision rule, we obtain:

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

This approach is a generalization of the source-channel approach. It has the advantage that additional models or feature functions can be easily integrated into the overall system. The model scaling factors λ_1^M are trained according to the maximum entropy principle, e.g. using the GIS algorithm. Alternatively, one can train them with respect to the final translation quality measured by some error criterion [4]. The overall architecture of the log-linear model combination is summarized in Figure 1.

5.3.1 The Alignment Template Approach

In this section, we give a brief description of the translation system, namely the alignment template approach. The key elements of this translation approach [3] are the *alignment templates*. These are pairs of source and target language phrases with an alignment within the phrases. The alignment templates are built at the level of word classes. This improves the generalization capability of the alignment templates.

The alignment templates are extracted from a word-aligned bilingual training corpus. We use GIZA++ to train this word alignment. The training procedure consists of five iterations with IBM model 1, five iterations with the HMM alignment model and three iterations with

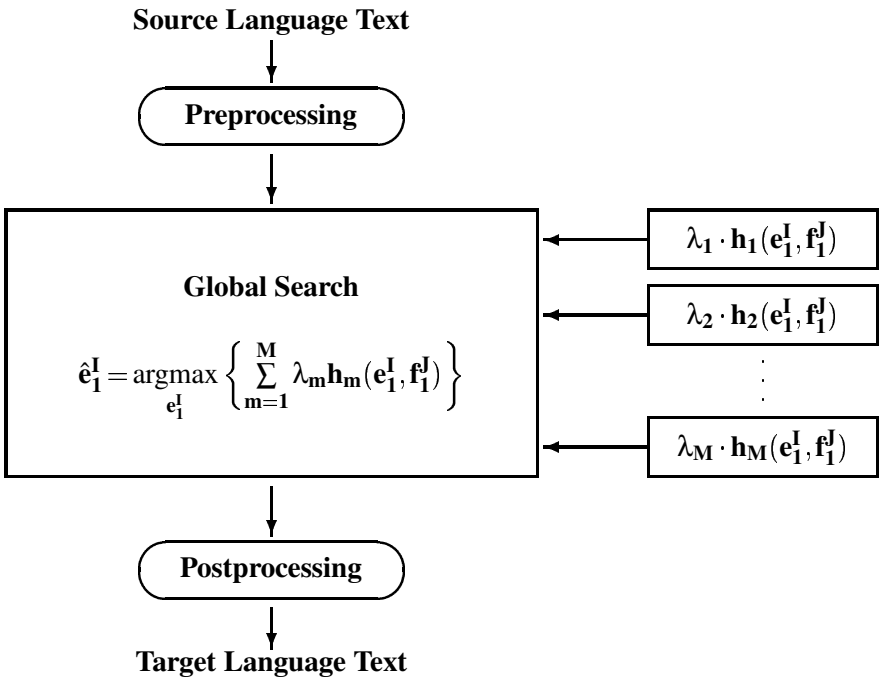

Figure 1: Architecture of the translation approach based on log-linear model combination.

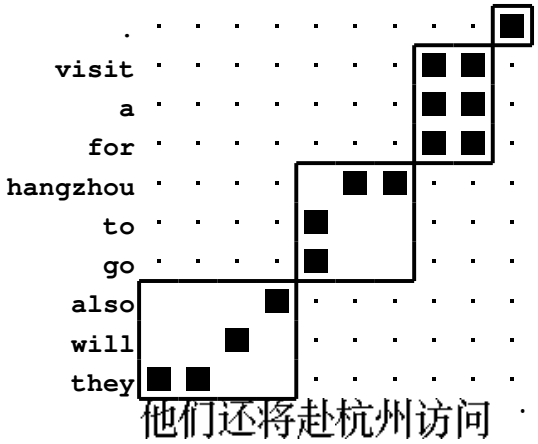

Figure 2: Example of a word-aligned sentence pair and some possible alignment templates.

IBM model 4. To obtain a more symmetric word alignment, we perform the training for both translation directions and unify the resulting Viterbi alignments.

Figure 2 shows an example of a word-aligned sentence pair. The word alignment is represented with the black boxes. The figure also includes some of the possible alignment templates, represented as the larger, unfilled rectangles. Note that the extraction algorithm would extract many more alignment templates from this sentence pair. In this example, the system input was the sequence of Chinese characters without any word segmentation.

A dynamic programming beam search algorithm is used to generate the translation hypothesis with maximum probability. For the Chinese–English, we do not permit reorderings of the

alignment templates. This means the search is monotone at the phrase level. Within the alignment templates, the reordering is learned in training and kept fix during the search process. There are no constraints on the reorderings within the alignment templates.

As mentioned before, we use a log-linear combination of various models: a phrase translation model as well as a word translation model. Additionally, we use two language models: a word-based trigram model and a class-based five-gram model. Furthermore, we use two heuristics, namely the word penalty and the alignment template penalty. The word penalty is a constant cost per produced target language word. It is a simple but quite effective way to adjust the translation hypotheses lengths. Using the alignment template penalty, which is a constant cost per used alignment template, it is possible to prefer longer alignment templates.

The models mentioned so far are integrated into the search algorithm. For some models this integration is not possible in an efficient way. In general, these are models that cannot be factorized along the target sentence positions. The IBM model 1 is one example as the whole target language sentence is necessary to compute the model probability. To integrate such models in our system, we use N-best list rescoring. So, our search algorithm generates a word graph of the most likely translation hypotheses. Out of this word graph we extract the N best translation candidates and compute for each of them additional model scores. For this evaluation, we performed rescoring with IBM model 1 and an additional language model. The model scaling factors are optimized with respect to the final translation quality measured with the BLEU score [5].

5.3.2 Data Resources

To train our statistical models, we make use of the following bilingual corpora provided by LDC: FBIS data, Hong Kong News Parallel Text, Hong Kong Hansards Parallel Text, English Translation of Chinese Treebank, Xinhua Chinese–English Parallel Text, Chinese–English Translation Lexicon, Sinorama Chinese–English Parallel Text, Chinese Treebank English Parallel Corpus, Chinese News Translation Corpus.

This data is preprocessed in the following way. The Chinese part is segmented using the LDC segmentation tool. The English part is tokenized and case information is removed, i.e. the corpus is converted to lowercase. We remove long sentences with more than 100 words from the training corpus. A rule-based categorization and translation of number and date expressions is performed. During the training procedure, number and date expressions are replaced with special symbols. During the translation process, the rule-based translation of the actual number or date is inserted via the alignment information.

During the post-processing we restore the case-information (true-casing) and do some text normalization. Our true-case mapper uses a maximum entropy tagger to distinguish the following five tags: all lowercase, initial letter uppercase, non-initial letter uppercase, all uppercase, not a word. This maximum entropy tagger is trained on some part of the English training corpus. The following features are used to classify the (lowercase) words into the five classes: word suffixes and word prefixes, the local context, i.e. words within a window of ± 2 words. The text normalization that is done during the post-processing includes removing double commas etc. and unifying monetary amounts, abbreviations and so on.

After the preprocessing, our training corpus consists of about three million sentences with somewhat more than 50 million running words. The corpus statistics of the preprocessed training corpus are shown in Table 4.

To train the language model, we use the English part of the bilingual training corpus and in addition we use the Xinhua News part of the English GigaWord corpus. The preprocessing of this additional monolingual data is the same as for the English part of the bilingual training

Table 4: Corpus statistics of the bilingual training data.

		Chinese	English
Train	Sentences	3.2M	
	Running Words	51.4M	55.5M
	Vocabulary	80 010	170 758
Lexicon	Entries	81 968	
Dev	Sentences	878	
	Running Words	26 431	23 694

corpus. This additional language model training data consists of about 155 million running words.

5.3.3 Summary

We use the alignment template approach which is a phrase-based translation model. The key idea is to memorize all phrasal translations that have been observed in the training corpus. We use a log-linear combination of various models. This allows the optimization of the model scaling factors with respect to the final evaluation criterion. Additional models, for example IBM model 1, can be easily integrated via rescoring of N-best lists.

References

[1] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

[2] F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

[3] F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

[4] F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

5.4 UKA: System Description

The statistical machine translation system developed in the Interactive Systems Laboratories (ISL) uses phrase-to-phrase translations as the primary building blocks to capture local context information, leading to better lexical choice and more reliable local reordering. A new approach to extract phrase translation pairs from bilingual data has been developed, which is not using the Viterbi alignment, but is based on optimizing a constrained word-to-word alignment for the entire sentence pair [Vogel 2004].

5.4.1 Training Data and Preprocessing

The training data for the baseline system was selected from the specified LDC corpora to include sentences of up to 30 words. As a preprocessing step, the Chinese side is segmented using the LDC segmenter, and a rule-based translation of number and date expressions is performed. The English side is tokenized and converted to lowercase.

5.4.2 The Models

Phrase Alignment The ISL translation system uses word-to-word and phrase-to-phrase translations, extracted from the bilingual corpus. Different phrase alignment methods have been explored in the past, like extracting phrase translation pairs from the Viterbi path of a word alignment, or simultaneously splitting source and target sentence into phrases and aligning them in an integrated way [Zhang 2003].

Phrase Alignment via Constrained Sentence Alignment Assume we are searching for a good translation for one source phrase $\tilde{f} = f_1 \ldots f_k$, and that we find a sentence in the bilingual corpus, which contains this phrase. We are now interested in finding a sequence of words $\tilde{e} = e_1 \ldots e_l$ in the target sentence, which is an optimal translation of the source phrase. Any sequence of words in the target sentence is a translation candidate, but most of them will not be considered translations of the source phrase at all, whereas some can be considered as partially correct translations, and a small number of candidates will be considered acceptable or good translations. We want to find these good candidates.

The IBM1 word alignment model aligns each source word to all target words with varying probabilities. Typically, only one or two words will have a high alignment probability, which for the IBM1 model is just the lexicon probability. We now modify the IBM1 alignment model by not summing the lexicon probabilities of all target words, but by restricting this summation in the following way:

- for words inside the source phrase we sum only over the probabilities for words inside the target phrase candidate, and for words outside of the source phrase we sum only over the probabilities for the words outside the target phrase candidates;
- the position alignment probability, which for the standard IBM1 alignment is $1/I$, where I is the number of words in the target sentence, is modified to $1/(l)$ inside the source phrase and to $1/(I-l)$ outside the source phrase.

More formally, we calculate the constrained alignment probability:

$$p_{i_1,i_2}(f|e) = \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} p(f_j|e_i) \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} p(f_j|e_i) \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} p(f_j|e_i)$$

and optimize over the target side boundaries i_1 and i_2.

$$(i_1, i_2) = \underset{i1,i2}{\operatorname{argmax}} \{p_{i_1,i_2}(f|e)\}$$

It is well know that 'looking from both sides' is better than calculating the alignment only in one direction, as the word alignment models are asymmetric with respect to aligning one to many words.

Similar to $p_{i_1,i_2}(f|e)$ we can calculate $p_{i_1,i_2}(e|f)$, now summing over the source words and multiplying along the target words:

$$p_{i_1,i_2}(e|f) = \prod_{i=1}^{i_1-1} \sum_{j \notin (j_1 \ldots j_2)} p(e_i|f_j) \times \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} p(e_i|f_j) \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1 \ldots j_2)} p(e_i|f_j)$$

To find the optimal target phrase we interpolate both alignment probabilities and take the pair (i_1, i_2) which gives the highest probability.

$$(i_1, i_2) = \underset{i_1,i_2}{\operatorname{argmax}} \{(1-c)p_{(i_1,i_2)}(f|e) + cp_{(i_1,i_2)}(f|e)\}$$

Actually, we take not only the best translation candidate, but all candidates which are within a given margin to the best one. All candidates are then used in the decoder, when also the language model is available to score the translations. The phrase pairs can be either extracted from the bilingual corpus at decoding time or stored and reused during system tuning. Single source words are treated in the same way, i.e. just as phrases of length 1. The target translation can then be one or several words.

Phrase Translation Probabilities Most phrase pairs $(\tilde{f}, \tilde{e}) = (f_{j_1} \ldots f_{j_2}, e_{i_1} \ldots e_{i_2})$ are seen only a few times, even in very large corpora. Therefore, probabilities based on occurrence counts have little discriminative power. In our system we calculate phrase translation probabilities based on a statistical lexicon, i.e. on the word translation probabilities $(p(f,e)$:

$$p(\tilde{f}|\tilde{e}) = \prod_j \sum_i p(f_j|e_i).$$

The Language Model The language model used in the decoder is a standard 3-gram language model. We use the SRI language model toolkit [SRI-LM Toolkit] to build language models of different sizes, using the target side of the bilingual data only or using additional monolingual data.

5.4.3 Decoding

The decoding process works in two stages: First, the word-to-word and phrase-to-phrase translations are used to generate a translation lattice. Second, a first-best or n-best search is performed on this lattice, using the language model probabilities in addition to the translation model probabilities to find the overall best translation.

Once the complete translation lattice has been built, a first-best search through this lattice is performed. In addition to the translation probabilities, or rather translation costs, as we use the negative logarithms of the probabilities for numerical stability, the language model costs are added and the path which minimizes the combined cost is returned.

Starting with a special begin-of-sentence hypothesis attached to the first node in the translation lattice, hypotheses are expanded over all outgoing edges from the current node. To realize word reordering, the search algorithm allows to leave a gap and jump to a distant node in the translation lattice, filling the gap at a later time. This requires to keep track of positions already covered in the source sentence.

The search space, especially when allowing for reordering, is very large. Pruning is applied to keep decoding times within reasonable bounds. Our decoder realizes a standard beam search, where all hypotheses which are worse than the best hypothesis by some factor are deleted [Vogel 2003].

5.4.4 Postprocessing

For the purposes of this evaluation, post-processing consisted of removing untranslated words, removing all punctuation marks for the ”plain transcription” conditions, and adding case information for the ”rich transcription” conditions. Case information was obtained by treating casing as a translation problem itself, training translation models on lower case/mixed case bi-text, and ”translating” using our decoder with word reordering disabled.

References

[Vogel 2004] Stephan Vogel, Sanjika Hewavitharana, Muntsin Kolss, Alex Waibel. The ISL Statistical Machine Translation System for Spoken Language Translation. *Int. Workshop on Spoken Language Translation*, 2004, Kyoto, Japan.

[SRI-LM Toolkit] SRILM - The SRI Language Modeling Toolkit. SRI Speech Technology and Research Laboratory. http://www.speech.sri.com/ projects/srilm/

[Vogel 2003] Stephan Vogel. SMT Decoder Disected: Word Reordering *Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003, Beijing, China.

[Zhang 2003] Ying Zhang, Stephan Vogel and Alex Waibel. Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation. *Proc. of International Confrerence on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003, Beijing, China.

4

A Source Language Format for SLT Baselines

```
<srcset setid="tc-star_VOM" srclang="Chinese">
<DOC docid="VOM19981003_0800_2179">
<p>
<seg id=1> unsegmented-chinese-text </seg>
...
<seg id=6> unsegmented-chinese-text </seg>
</p>
...
<seg id=19>  unsegmented-chinese-text </seg>
...
<seg id=30> unsegmented-chinese-text </seg>
</p>
</DOC>
...
<DOC docid="VOM19981007_1000_3345">
<p>
<seg id=1> unsegmented-chinese-text </seg>
...
<seg id=6> unsegmented-chinese-text </seg>
</p>
...
<seg id=12>  unsegmented-chinese-text </seg>
...
<seg id=17> unsegmented-chinese-text </seg>
</p>
</DOC>
</srcset>
```

B Target Language Format for SLT Baselines

```
<tstset setid="tc-star_VOM" srclang="Chinese" trglang="English">
<DOC docid="VOM19981003_0800_2179" sysid="lab.condition">
<p>
<seg id=1> english-translation </seg>
...
<seg id=6> english-translation </seg>
</p>
...
<seg id=19> english-translation </seg>
...
<seg id=30>  english-translation </seg>
</p>
</DOC>
...
<DOC docid="VOM19981007_1000_3345" sysid="lab.condition">
<p>
<seg id=1>  english-translation </seg>
```

```
...
<seg id=6>  english-translation </seg>
</p>
...
<seg id=12>  english-translation </seg>
...
<seg id=17>  english-translation </seg>
</p>
</DOC>
</tstset>
```

C List of Language Resource for Baseline Training.

No restrictions apply for monolingual resources.

The following list of permitted bilingual resources corresponds to the one defined for the Chinese-English Large Data Condition track of the 2004 NIST MT evaluation campaign.
Notice that one of the data sets, namely LDC2002T01, contains the VOA Mandarin transcripts used in the test data. Hence, participants are required to not use VOA transcripts contained in LDC2002T01.

LDC Code	Name
LDC2003E14	FBIS data
LDC2000T47	Hong Kong Laws Parallel Text
LDC2003E25	Hong Kong News Parallel Text,sentence-aligned
LDC2000T46	Hong Kong News Parallel Text
LDC2000T50	Hong Kong Hansard Parallel Text,aligned doc level
LDC2004E09	Hong Kong Hansard Parallel Text,aligned sent. level
LDC2002E17	English Translation of Chinese Treebank
LDC2002E18	Xinhua Chinese-English Parallel News Text v. 1.0 beta 2
LDC2004E12	UN Chinese-English Parallel Text Version 2
LDC2002L27	Chinese English Translation Lexicon version 3.0
LDC2002E58	Sinorama Chinese-English Parallel Text
LDC2002T01	Multiple-Translation Chinese Corpus (*)
LDC2003T17	Multiple-Translation Chinese Part 2
LDC2003E01	Chinese-English Name Entity Lists version 1.0 beta
LDC2003E04	Multiple Translation Chinese Corpus Part 3
LDC2004T05	Chinese Treebank Version 4.0
LDC2003E07	Chinese Treebank English Parallel Corpus
LDC2003E08	Chinese News Translation Corpus Part 1

(*) with exclusion of Voice of America Mandarin transcripts