

The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluation

Muntsin Kolss*, Bing Zhao†, Stephan Vogel†, Almut Silja Hildebrand†, Jan Niehues*,
Ashish Venugopal†, Ying Zhang†

*Institut für Theoretische Informatik
Universität Karlsruhe (TH), Karlsruhe, Germany
{kolss, jniehues}@ira.uka.de

† Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA, USA
{bzhao, vogel+, silja, ashishv, joy}@cs.cmu.edu

Abstract

In this paper we describe the ISL statistical machine translation system used in the TC-STAR Spring 2006 Evaluation campaign. This system is based on PESA phrase-to-phrase translations which are extracted from a bilingual corpus. The translation model, language model and other features are combined in a log-linear model during decoding. We participated in the Spanish Parliament (Cortes) and European Parliament Plenary Sessions (EPPS) task, in both the Spanish-to-English and English-to-Spanish direction, as well as the Chinese-to-English Broadcast News task, working on text input, manual transcriptions, and ASR input.

1. Introduction

TC-STAR - Technology and Corpora for Speech to Speech Translation is a three year integrated project financed by the European Commission within the Sixth Framework Programme. The aim of TC-STAR is to advance research in all core technologies for speech-to-speech translation (SST) in order to reduce the gap in performance between machines and human translators. To foster significant advances in all SST technologies, periodic competitive evaluations are conducted within TC-STAR for all components involved, including spoken language translation (SLT), as well as end-to-end systems.

Starting with the IBM system (Brown et al., 1993) in early 90's, statistical machine translation (SMT) has been the most promising approach for machine translation. Many approaches for SMT have been proposed since then (Wang and Waibel, 1998), (Och and Ney, 2000), (Yamada and Knight, 2001). Whereas the original IBM system was based on purely word translation models, current SMT systems incorporate more sophisticated models.

The ISL statistical machine translation system 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. In section 2., we describe the phrase alignment approach used by our system. Section 3. outlines the architecture of the decoder that combines the translation model, language model, and other models to generate the complete translation.

In section 4. we give an overview of the data and tasks and present evaluation results on the European Parliament Plenary Sessions (EPPS) task and the Chinese-to-English Broadcast News task.

2. Phrase Alignment

In this evaluation, we applied both the *phrase extraction via sentence alignment (PESA)* approach (Vogel, 2005) and a variation of the alignment-free approach, which is an extension to the previous work to extract bilingual phrase

pairs (Zhao and Vogel, 2005). In the extended system, we used eleven feature functions including phrase level IBM Model-1 probabilities and phrase level fertilities to locate the phrase pairs from the parallel training sentence pairs. The feature functions are then combined in a log-linear model as follows:

$$P(X|e, f) = \frac{\exp(\sum_{m=1}^M \lambda_m \phi_m(X, e, f))}{\sum_{\{X'\}} \exp(\sum_{m=1}^M \lambda_m \phi_m(X', e, f))}$$

where $X \rightarrow (f_j^{j+l}, e_i^{i+k})$ corresponds to a phrase-pair candidate extracted from a given sentence-pair (e, f) ; ϕ_m is a feature function designed to be informative for phrase extraction. Feature function weights $\{\lambda_m\}$, are the same as in our previous experiments (Zhao and Waibel, 2005). This log-linear model serves as a performance measure function in a local search. The search starts from fetching a test-set specific source phrase (e.g. Chinese ngram); it localizes the candidate ngram's center in the English sentence; and then around the projected center, it finds out all the candidate phrase pairs ranked with the log-linear model scores. In the local search, down-hill moves are allowed so that functional words can be attached to the left or right boundaries of the candidate phrase-pairs.

The feature functions that compute different aspects of phrase pair (f_j^{j+l}, e_i^{i+k}) are as follows:

- Four compute the IBM Model-1 scores for the phrase-pairs $P(f_j^{j+l}|e_i^{i+k})$ and $P(e_i^{i+k}|f_j^{j+l})$; the remaining parts of (e, f) excluding the phrase-pair is modeled by $P(f_{j' \notin [j, j+l]}|e_{i' \notin [i, i+k]})$ and $P(e_{i' \notin [i, i+k]}|f_{j' \notin [j, j+l]})$ using the translation lexicons of $P(f|e)$ and $P(e|f)$.
- Another four of them compute the phrase-level length relevance: $P(l+1|e_i^{i+k})$ and $P(J-l-1|e_{i' \notin [i, i+k]})$, where $e_{i' \notin [i, i+k]}$ is denoted as the remaining English words in e : $e_{i' \notin [i, i+k]} = \{e_{i'} | i' \notin [i, i+k]\}$, and J is the length of f . The probability is computed via dynamic programming using the English

word-fertility table $P(\phi|e_i)$. $P(k+1|f_j^{j+l})$ and $P(I-k-1|f_{j' \notin [j, j+l]})$ are computed in a similar way.

- Another two of the scores aim to bracket the sentence pair with the phrase-pair as detailed in (Zhao and Waibel, 2005).
- The last function computes the average word alignment links per source word in the candidate phrase-pair.

We assume each phrase-pair should contain at least one word alignment link. We train the IBM Model-4 with GIZA++ (Och and Ney, 2003) in both directions and grow the intersection with word pairs in the union to collect the word alignment. Because of the last feature-function, our approach is no longer truly “alignment-free”. More details of the log-linear model and experimental analysis of the feature-functions are given in (Zhao and Waibel, 2005).

When using the extracted phrase pairs for translating a test sentence, a slightly different set of features is used as translation model score. In the extended system, we pass eight scores to the decoder: Relative phrase frequencies in both directions, phrase-level fertility scores for both directions computed via dynamic programming, the standard IBM Model-1 scores for both directions (i.e. $P(f_j^{j+l}|e_i^{i+k}) = \prod_{j' \in [j, j+l]} \sum_{i' \in [i, i+k]} P(f_{j'}|e_{i'}) / (k+1)$), and the unnormalized IBM Model-1 scores for both directions (i.e. $P(f_j^{j+l}|e_i^{i+k}) = \prod_{j' \in [j, j+l]} \sum_{i' \in [i, i+k]} P(f_{j'}|e_{i'})$).

The individual scores are then combined via the optimization component of the decoder (e.g. Max-BLEU optimization) as described in section 3. in the hope of balancing the sentence length penalty.

2.1. Integrated Sentence Splitting

The underlying statistical word-to-word alignment used for phrase alignment can in principle be based on any statistical word-to-word alignment. In this evaluation, IBM Model-1 trained in both directions was used exclusively for the Spanish-to-English and English-to-Spanish translation directions, as using higher IBM models did not improve the phrase alignment quality on the respective development sets. For the Chinese-to-English Broadcast News task, IBM Model-4 trained with GIZA++ (Och and Ney, 2003) was used.

For IBM Model-1, a small improvement came from splitting long training sentences during lexicon training, similar to the method described in (Xu et al., 2005). Splitting long sentences improves training time as well as lexicon perplexity. In our scheme, potential split points are defined in both source and target training sentences at parallel punctuation marks. Each of these punctuation marks produces a three-way split, with the punctuation mark forming the middle sentence part.

Training sentence pairs are split iteratively by the following procedure to choose split points: We calculate the lexicon probability of the un-split sentence pair as well as of the left, right and middle partial sentence pairs, and re-calculate the lexicon and split the best N sentence pairs, in each iteration, until a predefined maximal sentence length or maximal number of splits has been reached.

The actual phrase alignment is then performed on the original, un-split training corpus.

3. Decoder

The beam search decoder combines all model scores to find the best translation. In the TC-STAR evaluation, the following models were used:

- The translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus, annotated with multiple translation model scores, as described in section 2..
- A trigram language model. The SRI language model toolkit was used to train the models (Technology and Laboratory,). Modified Kneser-Ney smoothing was used throughout.
- A word reordering model, which assigns higher costs to longer distance reordering. We replace the jump probabilities $p(j|j', J)$ of the HMM word alignment model

$$p(j|j', J) = \frac{\text{count}_J(j - j')}{\sum_{j''} \text{count}_J(j - j'')}$$

by a simple Gaussian distribution:

$$p(j|j', J) = e^{-|j-j'|}$$

where j is the current position in the source sentence, j' is the previous position, and J is the number of words in the source sentence.

- Simple word and phrase count models. The former is essentially used to compensate for the tendency of the language model to prefer shorter translations, while the latter can give preference to longer phrases, potentially improving fluency.

The decoding process is organized into two stages:

- Find all available word and phrase translations. These are inserted into a lattice structure, called translation lattice.
- Find the best combination of these partial translations, such that every word in the source sentence is covered exactly once. This amounts to doing a best path search through the translation lattice, which is extended to allow for word reordering.

In addition, the system needs to be optimized. For each model used in the decoder a scaling factor can be used to modify the contribution of this model to the overall score. Varying these scaling factors can change the performance of the system considerably. Minimum error training is used to find a good set of scaling factors.

In the following sub-sections, these different steps will be described in some more detail.

3.1. Building A Translation Lattice

The ISL SMT decoder can use phrase tables, generated at training time, but can also do just-in-time phrase alignment. This means that the entire bilingual corpus is loaded and the source side indexed using a suffix array (Zhang and Vogel, 2005). For all ngrams in the test sentence, occurrences in the corpus are located using the suffix array. For a number of occurrences, where the number can be given as a parameter to the decoder, phrase alignment as described in section 2. is performed and the found target phrase added to the translation lattice.

If phrase translations have already been collected during training time, then this phrase table is loaded into the decoder and a prefix tree constructed over the source phrases. This is typically done for high-frequency source phrases and allows for an efficient search to find all source phrases in the phrase table which match a sequence of words in the test sentence. If a source phrase is found in the phrase translation table then a new edge is added to the translation lattice for each translation associated with the source phrase.

Each edge carries not only the target phrase, but also a number of model scores. There can be several phrase translation model scores, calculated from relative frequency, word lexicon and word fertility. In addition, the sentence stretch model score and the phrase length model score are applied at this stage.

3.2. Searching for the Best Path

The second stage in the decoding is finding a best path through the translation lattice. 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.

To search for the best translation means to generate partial translations, i.e. a sequence of target language words which are translations of some of the source words, and a score. These hypotheses are expanded into longer translations until the entire source sentence has been accounted for. To restrict the search space, only limited word reordering is done. Essentially, decoding runs from left to right over the source sentence, but words can be skipped within a restricted reordering window and translated later. In other words, the difference between the highest index of already translated words and the index of still untranslated words is smaller than a specified constant, which typically is 4.

When a hypothesis is expanded, the language model is applied to all target words attached to the edge over which the hypothesis is expanded. In addition, the distortion model is applied, adding a cost depending on the distance of the jump made in the source sentence.

Hypotheses are recombined whenever the models can not change the ranking of alternative hypotheses in the future. For example, when using a trigram language model, two hypotheses having the same two words at the end of the word sequences generated so far, will get the same increment in language model scores when expanded with an additional word. Therefore, only the better hypothesis needs to be expanded. The translation model and distor-

tion model require that only the hypotheses which cover the same source words are compared. In addition to total source side coverage, the decoder can optionally use the language model history and the target sentence length to distinguish hypotheses.

		Spanish	English
Training	Sentences	1,242,811	
	Words	30,554,408	29,579,969
	Vocabulary	126,300	80,535
Es-En FTE Test	Sentences	1,782	
	Words	56,596	-
	Vocabulary	6,713	-
	Unknown	229	-
Es-En Verbatim Test	Sentences	1,596	
	Words	61,227	-
	Vocabulary	6,674	-
	Unknown	245	-
Es-En ASR Test	Sentences	2,225	
	Words	61,174	-
	Vocabulary	6,848	-
	Unknown	73	-
En-Es FTE Test	Sentences	1,117	
	Words	-	28,494
	Vocabulary	-	3,897
	Unknown	-	71
En-Es Verbatim Test	Sentences	1,155	
	Words	-	30,553
	Vocabulary	-	3,955
	Unknown	-	97
En-Es ASR Test	Sentences	893	
	Words	-	31,076
	Vocabulary	-	3,972
	Unknown	-	22

Table 1: *Corpus statistics for EPPS and Cortes.*

As typically too many hypotheses are generated, pruning is necessary. This means that low-scoring hypotheses are removed. Similar to selecting a set of features to decide when hypotheses can be recombined, a set of features is selected to decide when hypotheses are compared for pruning. By dropping one or two of the criteria for recombination, a mapping of all hypotheses into a number of equivalence classes is created. Within each equivalence class, only hypotheses which are close to the best one are kept. Pruning can be done with more equivalence classes and smaller beam, or coarser equivalence classes and wider beams. For example, comparing all hypotheses, which have translated the same number of source words, no matter what the final two words are, would be working with a small number of equivalence classes in pruning.

		Chinese	English
Training	Sentences	22,137,200	
	Words	200,076,900	212,814,379
	Vocabulary	232,434	505,397
Verbatim Test	Sentences	1,232	
	Words	29,889	-
	Vocabulary	4,782	-
	Unknown	26	-
ASR Test	Sentences	1,286	
	Words	32,786	-
	Vocabulary	5,085	-
	Unknown	27	-

Table 2: *Corpus statistics for Chinese-English.*

3.3. Optimizing the system

Each model contributes to the total score of the translation hypotheses. As these models are only approximations to the real phenomena they are supposed to describe, and as they are trained on varying, but always limited data, their reliability is restricted. However, the reliability of one model might be higher than the reliability of another model. So, we should put more weight on this model in the overall decision. This can be done by doing a log-linear combination of the models. In other words, each model score is weighted and we have to find an optimal set of these weights or scaling factors. When dealing with two or three models, grid search is still feasible. When adding more and more features (models) this no longer is the case and automatic optimization needs to be done. We use the Minimum Error Training as described in (Och, 2003), which uses rescoring of the n-best list to find the scaling factors with maximize BLEU or NIST score.

Starting with some reasonably chosen model weights a first decoding for some development test set is done. An n-best list is generated, typically a 1000-best list. Then a multi-linear search is performed, for each model weight in turn. The weight, for which the change gives the best improvement in the MT evaluation metric, is then fixed to the new value, and the search repeated, till no further improvement is possible.

The optimization is therefore based on an n-best list, which resulted from sub-optimal model weights, and contained only a limited number of alternative translations. To eliminate any restricting effect, a new full translation is done with the new model weights. The resulting new n-best list is then merged to the old n-best list, and the entire optimization process repeated. Typically, after three iterations of doing translation plus optimization, translation quality, as measured by the MT evaluation metric, converges. More details on our optimization procedure are found in (Venugopal et al., 2005) and (Venugopal and Vogel, 2005).

4. Evaluation

For spoken language translation, the TC-STAR Spring 2006 evaluation consisted of two main tasks: Mandarin

Chinese to English translation of broadcast news, recorded from *Voice of America* radio shows, and translation of parliamentary speeches from Spanish to English, and from English to Spanish. The parliamentary speech data was taken from native speakers in the European Parliament (*EPPS* subtask), and, in the case of Spanish, partly from the Spanish National Parliament (*Cortes* subtask). For each of the three translation directions, there were multiple input conditions: *ASR input*, consisting of speech recognizer output provided by TC-STAR ASR partners, *verbatim input*, the manual transcriptions of the audio data, and *FTE input* for the parliamentary speech tasks, the edited final text edition of the parliamentary sessions published on the parliament’s website. We participated in all translation directions and all input conditions, in the primary track, i.e., using no additional training data other than specified.

We report translation results using the well known evaluation metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002), as well as WER and PER. All measures reported here were calculated using case-sensitive scoring on two reference translations per test set. The method described in (Matusov et al., 2005) was used to score the automatically segmented ASR input test sets.

4.1. Spanish-English and English-Spanish EPPS and Cortes task

The Spanish-to-English and English-to-Spanish evaluation systems were trained on the supplied symmetrical EPPS training corpus. As a preprocessing step, we separated punctuation marks from words in both source and target side and converted the text into lowercase. Sentence pairs differing in length by a factor of more than 1.5 were discarded.

The same preprocessing was applied to test sets. For the ASR test sets, we used the automatic segmentation defined by punctuation marks to separate the test data into sentence-like units. Table 1 shows the training and test corpus statistics after preprocessing.

For scoring, generated punctuation marks were re-attached to words, and a truecasing module was run to restore case information. Our truecasing module treats case restoration as a translation problem, using a simple translation model base on relative frequencies of true-cased words and a case-sensitive target language trigram language model, trained on the appropriate training corpus side.

Table 3 summarizes the official translation results for our primary submissions. Not surprisingly, MT scores on ASR hypotheses were lower than on text and verbatim transcriptions, due to ASR word error rates on the input side of 6.9% for English and 8.1% overall for Spanish (14.5% and 9.5% respectively when including punctuation marks).

Example translation output from the Spanish-to-English system is shown in table 5.

4.2. Chinese-English Broadcast News task

Parallel training data for the Chinese-to-English evaluation system consisted of about 200 million words for each language, taken from the LDC corpora *FBIS*, *UN Chinese-English Parallel Text*, *Hong Kong Parallel Text*, and *Xinhua Chinese-English Parallel News Text*.

Task	Input condition	Translation Direction	NIST	Bleu [%]	WER [%]	PER [%]
EPPS	FTE	Spanish-English	10.60	52.3	37.0	27.1
EPPS	Verbatim	Spanish-English	9.85	46.0	43.2	30.3
EPPS	ASR	Spanish-English	8.53	33.0	52.2	39.3
Cortes	FTE	Spanish-English	8.85	39.0	50.0	36.0
Cortes	Verbatim	Spanish-English	8.84	38.1	51.1	35.2
Cortes	ASR	Spanish-English	7.18	24.6	63.7	46.5
EPPS	FTE	English-Spanish	9.56	44.0	43.6	33.7
EPPS	Verbatim	English-Spanish	9.08	40.1	47.6	36.1
EPPS	ASR	English-Spanish	8.10	31.3	55.5	43.1

Table 3: EPPS and Cortes Task: Official results for the primary submissions.

Input condition	Translation Direction	NIST	Bleu [%]	WER [%]	PER [%]
Verbatim	Chinese-English	5.48	10.8	81.9	61.7
ASR	Chinese-English	4.59	08.5	82.9	66.9

Table 4: Chinese-English Broadcast News Task: Official results for the primary submissions.

Pre- and postprocessing on the English side was similar to that for Spanish and English. For Chinese, preprocessing included re-segmenting Chinese characters into words using the LDC segmenter, and a limited amount of rule-based translation of number and date expressions. Table 2 shows the training and test corpus statistics after preprocessing. The official translation results for our primary submissions are summarized in table 4. For the ASR input condition, the ASR character error rate was 9.8%, leading to the observed drop in MT scores.

5. Conclusion

In this paper we described the ISL statistical machine translation system that was used in the TC-STAR Spring 2006 Evaluation campaign. Our system, built around extraction of PESA phrase-to-phrase translation pairs, was applied to all translation directions and input conditions.

A brief analysis shows that further work is needed to bring translation performance on Chinese-to-English Broadcast News up to the level that is available today for translating between Spanish and English parliamentary speeches.

6. Acknowledgements

This work has been funded by the European Union under the integrated project TC-Star - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, <http://www.tc-star.org>).

7. References

- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *In Proceedings of the Human Language Technology Conference (HLT)*, San Diego, CA, March.
- E. Matusov, G. Leusch, O. Bender, and H. Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *In Proceedings of the International Workshop Spoken Language Translation (IWSLT)*, Pittsburgh, PA, October.
- Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Prof. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.
- Speech Technology and Research Laboratory. The sri language modeling toolkit. <http://speech.sri.com/projects/srilm/>.
- Ashish Venugopal and Stephan Vogel. 2005. Considerations in maximum mutual information and minimum classification error training for statistical machine translation. In *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*, Budapest, Hungary, May.
- Ashish Venugopal, Andreas Zollmann, and Alex Waibel. 2005. Training and evaluation error minimization rules for statistical machine translation. In *Proceedings of ACL 2005, Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, MI, June.
- Stephan Vogel. 2005. PESA: Phrase pair extraction as sentence splitting. In *Proc. of the Machine Translation Summit X*, Phuket, Thailand, September.

Verbatim input	para Rumanía y Bulgaria la adhesión significará, como sabemos bien por experiencia propia, el mejor camino para la modernización, la estabilidad, el progreso y la democracia.
ASR input	para Rumanía y Bulgaria la emisión significará como estamos bien por experiencia propia el mejor camino para la modernización la estabilidad el progreso y la democracia .
Output on Verbatim	For Romania and Bulgaria accession will mean, as we well know from my own experience, the best way for the modernisation, stability, the progress and democracy.
Output on ASR	For Romania and Bulgaria the emission means as we are well by my own experience the best way for modernisation stability the progress and democracy.
Reference Text	For Rumania and Bulgaria, accession will mean, as we are well aware from our own experience, the best path towards modernisation, stability, progress and democracy.
Verbatim input	la ampliación constituye por ello una responsabilidad histórica , un deber de solidaridad y un proyecto político económico de primera magnitud para el futuro de Europa .
ASR input	la ampliación constituye por ello una responsabilidad histórica un deber de solidaridad y un proyecto político económico de primera magnitud para el futuro de Europa .
Output on Verbatim	Enlargement therefore constitutes a historic responsibility, a duty of solidarity and a political project of economic first magnitude for the future of Europe.
Output on ASR	Enlargement therefore constitutes an historic responsibility a duty of solidarity and a political project of economic first magnitude for the future of Europe.
Reference Text	Thus, the enlargement entails a historical responsibility, a duty of solidarity and a political-economic project of prime magnitude for Europe's future.

Table 5: Example translation output for Spanish-to-English.

- Yeyi Wang and Alex Waibel. 1998. Fast decoding for statistical machine translation. In *Proc. of the ICSLP 98*, pages 2775–2778, Sidney, Australia, December.
- J. Xu, R. Zens, and H. Ney. 2005. Sentence segmentation using ibm word alignment model 1. In *Proc. of the European Association for Machine Translation 10th Annual Conference (EAMT 2005)*, pages 280–287, Budapest, Hungary, May.
- Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 523–530, Toulouse, France, July.
- Ying Zhang and Stephan Vogel. 2005. Competitive grouping in integrated phrase segmentation and alignment model. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 159–162, Ann Arbor, Michigan, June.
- Bing Zhao and Stephan Vogel. 2005. A generalized alignment-free phrase extraction. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 141–144, Ann Arbor, Michigan, June.
- Bing Zhao and Alex Waibel. 2005. Learning a log-linear model with bilingual phrase-pair features for statistical machine translation. In *Proceedings of the SigHan Workshop*, Jeju, Korea, October.