Integrated N-best Re-ranking for Spoken Language Translation

V. H. Quan, M. Federico, M. Cettolo

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
via Sommarive, 18 - Povo di Trento - Italy
{vhquan,federico,cettolo}@itc.it

Abstract

This paper describes the application of N-best lists to a spoken language translation system. Multiple hypotheses are generated both by the speech recognizer and by the statistical machine translator; they are finally re-ranked by optimally weighting recognition and translation scores, estimated in an integrated scheme. We provide experimental results for the Italian-to-English direction on the BTEC corpus, a collection of sentences in the touristic domain developed within the C-STAR project.

1. Introduction

In comparison with written language, speech and especially spontaneous speech poses additional difficulties for the task of automatic translation. Typically, these difficulties are caused by errors of the speech recognition step, which is carried out before the translation process. As a result, the sentence to be translated is not necessarily well-formed from a syntactic point-of-view. Even without recognition errors, speech translation has to cope with a lack of conventional syntactic structures because structures of spontaneous speech differ from those of written language. Recently, the statistical approach for machine translation showed the potential to tackle these problems for the following reasons. First, the statistical approach is able to avoid hard decisions at any level of the translation process. Second, for any source sentence, a translated sentence in the target language is guaranteed to be generated. In most cases, this will be hopefully a syntactically perfect sentence in the target language; but even if this is not the case, in most cases, the translated sentence will convey the meaning of the spoken sentence [1].

Currently, statistical speech translation systems show typically a cascaded structure: speech recognition followed by machine translation. This structure lacks some joint optimality in performance since the speech recognition module and translation module are running rather independently. In fact, the translation module of a speech translation system usually takes a single best recognition hypothesis and performs standard 1-best text-based translation. Lots of supplementary information available from speech recognition such as N-best list, word graphs, confusion networks and likelihoods of acoustic (AM) and language model (LM) are typically not well utilized in the translation process. This kind of information can be effective for improving translation quality if employed properly [2, 3].

The main objective of this work is to measure the gain obtained by extending the baseline ITC-irst spoken language translation (SLT) system to the use of N-best lists both as input from the speech recognizer and as output toward a re-ranking module. We provide experimental results for the Italian-to-English direction on the BTEC corpus, a collection of sentences in the touristic domain. A statistical significant improvement of the BLEU score is measured with respect to the baseline system.

2. The ITC-irst SLT System

The ITC-irst statistical machine translation system [4] implements an extension of the IBM Model 4 as a log-linear interpolation of statistical models, which apply probabilities at the level of *phrases*. The interpolation involves the following models: lexicon, distortion, fertility and target LM. The use of phrases rather than words is a mean to cope with the limited context that Model 4 exploits to guess word translation (lexicon model) and word positions (distortion model).

The scaling factors of the log-linear model are estimated by the *minimum error training* procedure as described in [5].

Figure 1 illustrates the ITC-irst SLT systems, which can be virtually divided into two parts. In the left-hand side, beginning from the speech signal of the utterance, the automatic speech recognition (ASR) produces a word graph that contains alternative recognition hypotheses.

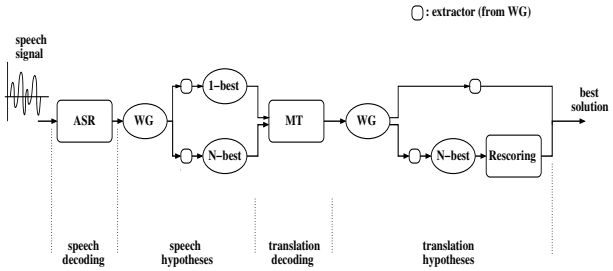

Figure 1: The ITC-irst Spoken Language Translation System.

By using the generated word graph, either the best hypothesis or the N-best list can be extracted and passed to the text machine translation module (MT). Similarly, in the right-hand side, the output of machine translation is again a word graph, which compactly encloses multiple translation hypotheses in the target language. The best translation hypothesis can be extracted directly from the word graph. Additionally, the possibility of having word graphs/N-best lists as output of the machine translation process allows to employ deeper and more extensive knowledge sources for rescoring them.

In this work, focus is given to the system highlighted in Figure 1, which handles multiple hypotheses in form of N-best lists both from the ASR and from the MT outputs. Since the two modules actually build word graphs, the algorithm presented in [6] was implemented for extracting N-best lists from them.

3. System Parameters Tuning

Both the ASR and the MT modules make use of weights to combine their statistical models. In particular, in the speech recognizer the LM scores are scaled by a factor which allows to get values comparable to those of the AM; on the other side, the translation decoder relies on the weighted log-linear interpolation of the lexicon, distortion (actually a pair of models), fertility (again a pair) and target LMs.

The ASR weight of the source LM and the six MT weights are optimized on a development set by minimizing an error function, as proposed in [7]. We implemented the iterative procedure described in [5], which uses the *simplex* algorithm.

In the case of ASR, the error function to be optimized is the word error rate (WER). Figure 2 shows the procedure that is employed for estimating the optimal parameters. By using some values $\{\lambda_1, \lambda_2\}$ at step t, the ASR module processes sentences of the development set. The set of 1-best hypotheses is compared to the reference texts in order to get the WER, which is the value the function to be optimized gets on $\lambda^{(t)}$'s. On the basis of this value, the simplex algorithm selects a new $\{\lambda_1, \lambda_2\}$ which goes toward a (local) minimum of the function. The procedure is iterated until a convergence criterion is met and eventually the optimal parameters are generated.

Actually, since the weight of the acoustic model can be kept fixed to 1, only the LM weight has to be estimated.

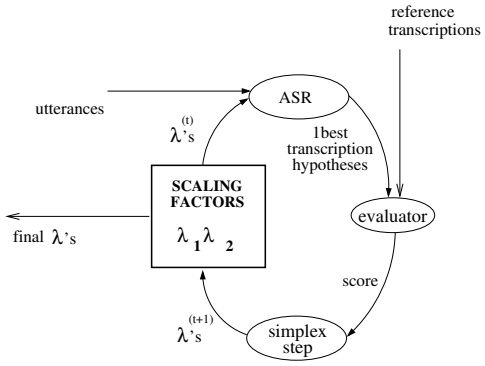

Figure 2: Estimation of ASR parameters.

The same procedure is applied for estimating the six MT parameters (Figure 3). In this case the error function to be optimized is the BLEU score [8].

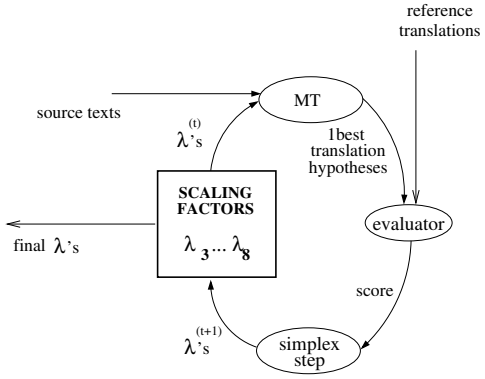

Figure 3: Estimation of MT parameters.

4. Re-Ranking Module

Once the parameters of the SLT system have been optimized, it is able to generate word graphs at its best. What is actually generated is the following: given an input sentence, the ASR module builds a word graph from which N-best transcription hypotheses are extracted. Each of them is passed to the MT module: from the generated word graph M-best translation hypotheses are extracted. Summarizing, for each input sentence, the SLT system generates NxM-best translation hypotheses.

Each entry of the NxM-best list is characterized by the 8 scores mentioned above, two coming from the ASR and the remaining six from the MT module. In [3], the re-ranking of the list by using score weights estimated on a development set by means of a minimum error rate procedure is proposed; this allows to integrate in a tight way ASR and MT features. Of course, it would be possible also add new features (scores) not employed during the processing [9].

In this work we have followed the integration approach suggested in [3]. Figure 4 illustrates the scheme for the estimation of optimal re-ranking weights.

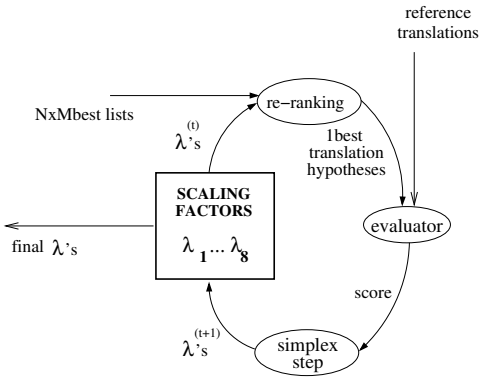

Figure 4: Estimation of weights for re-ranking

At each step of the procedure, the NxM-best lists are re-ranked by means of the λ's estimated in the previous step. With the set of the best translation hypotheses, the BLEU score is computed which is used by the simplex algorithm for choosing the new λ values. The loop ends when some ending criterion is met and the found optimal values of parameters are output.

5. Experimental Results

In order to investigate the impact of the techniques described in the previous sections on speech translation, experiments on the Basic Traveling Expressions Corpus (BTEC) has been conducted.

An updated version of the ASR system developed at ITC-irst and described in [10], and the MT decoder described in [4] upgraded accordingly to the contents of previous sections have been employed.

5.1. Training Data

The BTEC corpus, jointly developed by the partners of the C-STAR project[1], is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country. The initial collection of Japanese and English

[1] www.c-star.org

sentence pairs was translated into Chinese, Korean and Italian, as reported in [11]. Then, as we are interested in Italian-to-English translation, we selected pair of sentences in those two languages.

In Table 1 detailed statistics on the training set are reported. It was employed for the training of both the ASR source LM and the MT models. Estimation of MT models needs the availability of bidirectional alignments and estimates of Model 4, which were computed with the GIZA++ toolkit [12]. Translation phrase-pairs were extracted from the training corpus according to the method described in [4]. The number of extracted phrase-pairs was 1.1M.

	source		target		phrase
#sent.	W	|V|	W	|V|	pairs
52K	451K	15.7K	480K	10.8K	1074K

Table 1: The BTEC training set.

On the other side, the original AM used for broadcast news transcription has been kept, i.e. no specific training nor adaptation has been performed to the BTEC task. In fact, the ASR AM was trained on recordings of Italian radio and television news programs as described in [10], even if the currently employed model has been trained on much more data (130 hours in total). The additional corpus consists of audio recordings of television news programs, automatically transcribed by exploiting close-captions provided by the broadcaster.

5.2. Test and Development Sets

Table 2 gives detailed statistics about the development and test sets used in the experiments. Figures related to the target language refer to the gold reference, which is the only English reference used for evaluation purposes. The perplexity of the ASR LM on the test set is around 50.

		source		target			speech
set	#sent.	W	|V|	W	|V|	#spk	(m)
dev	500	3954	957	3979	765	5f/5m	34.0
test	3006	23409	2817	23990	2059	8f/9m	205.2

Table 2: The BTEC development and test sets.

5.3. ASR Tuning

The optimal LM weight on the development set was estimated according to the procedure described in Section 3.

Figure 5 draws the WER of the development set as a function of the LM weight. The value 9.25 yields the minimum WER. With that weight, performance on development and test sets are reported in Table 3.

set	WER	95% conf. interval
dev	20.93	18.76 - 23.02
test	22.41	20.19 - 24.80

Table 3: ASR performance with the optimal LM weight.

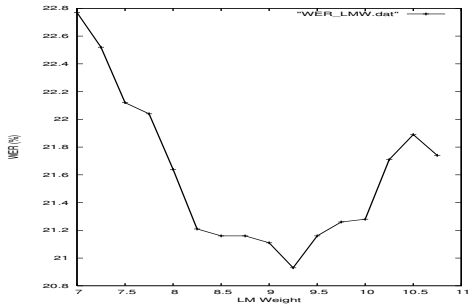

Figure 5: WER vs. LM weight on the development set.

5.4. MT Tuning

The set of optimal MT λ's was estimated by means of the scheme proposed in Section 3 on the development set. The simplex algorithm took 140 iterations to converge. Table 4 gives BLEU scores on both development and test sets by employing non-estimated and estimated λ's.

	λ's	
set	1 . . . 1	estimated
dev	50.93	54.07
test	52.63	53.15

Table 4: BLEU score for text translation using estimated and non-estimated λ's.

5.5. Re-ranking Weights Estimation

The integration of ASR and MT scores is performed by means of the re-ranking procedure described in Section 4. By using the optimal λ's for ASR and MT, a list of 100x100best was generated for each sentence of the development set. The set of these 100x100best lists was used for the estimation of re-ranking weights.

Table 5 gives BLEU scores on the development set by employing non-estimated and estimated λ's. In this case, non-estimated values were set to 1 for MT scores and to 0.05 and 0.5 for, respectively, AM and source LM scores. These values represent the starting λ's given to the simplex, which guarantee a quite good ratio between the AM and LM ASR scores (1:10 instead of the optimal 1:9.25) and rescale ASR scores in order to make their dynamic comparable to that of MT scores.

The simplex algorithm took 142 iterations to converge.

	λ's	
set	0.05 0.5 1 . . . 1	estimated
dev	37.92	40.11

Table 5: BLEU score for dev set 100x100best lists re-ranked by using estimated and non-estimated λ's.

5.6. Final SLT Results

Finally, the 100x100best lists generated for each sentence of the test set have been re-ranked by employing the re-ranking weights estimated on the development set. Table 6 gives BLEU scores after the final re-ranking step. For comparison purposes, performance of the following systems is given:

`baseline`: 1best transcription from the optimal ASR translated as 1best by the baseline MT (λ's equal to 1)

`optimal SLT`: 1best transcription from the optimal ASR translated as 1best by the optimal MT (estimated λ's)

For all scores, the 95% confidence interval is also provided.

system	BLEU	95% conf. interval
baseline	39.66	38.49 - 40.79
optimal SLT	40.02	38.88 - 41.18
optimal SLT + re-ranking	41.22	40.03 - 42.42

Table 6: Baseline and optimal SLT systems performance.

It is worth to noticing that the optimally tuned system with re-ranking gives a statistically significant improvement on the baseline performance, although no really new source of information has been employed. It also has to be highlighted the degradation of performance due to the use of an input corrupted by recognizer errors, as the BLEU score from around 53 (Table 4) decreases up to about 41 (Table 6).

6. Conclusions and Future Work

In this work the first attempt for improving the quality of the baseline speech translation system described in [5] has been presented. Specifically, by exploiting the word graph generation, we were able to produce NxM-best translation candidates as the output of the speech translation system, where N stands for multiple transcription hypotheses and M for multiple translation hypotheses. The NxM-best lists were used in a parameter tuning scheme whose final goal is to optimize the parameters of the SLT system, which includes a re-ranking module. On the BTEC corpus, a statistical significant improvement has been measured in the translation quality with respect to the performance of the baseline system.

Some issues which should further improve our SLT system will be investigated in the future:

`Additional features`. Currently, the re-ranking of NxM-best lists is performed on scores of models employed by the ASR and MT during the decoding. In addition to those 8 scores, other features could also be computed on each candidate in order to allow a more refined re-ranking, such as higher order and/or part-of speech LMs, penalty-length models, IBM model 1, jump weights, maximum entropy alignment models, (dynamic) example matching scores, etc.

`Confusion network MT decoder`. A special MT algorithm has been developed at ITC-irst for dealing with confusion networks as input. A confusion network is a word graph in which each path from the starting to the ending nodes passes through all nodes. By using confusion networks, multiple transcription hypotheses from the ASR can be processed in one shot, instead of performing N MT decodings as in the N-best list case.

`Translation rules`. Recently, the ITC-irst SLT decoder has also been made able to cope with explicit translation rules. This should improve the quality of the translation in general, since it will allow to force both verbatim translation of proper names and the correct translation of some specific patterns.

7. Acknowledgements

This work was partially financed by the European Commission under the project TC-STAR - Technology and Corpora for Speech to Speech Translation Research (IST-2002-2.3.1.6, http://www.tc-star.org).

8. References

[1] F. J. Och and H. Ney, "Improved Statistical Alignment Models," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, 2000.

[2] H. Ney, "Speech Translation: Coupling of Recognition and Translation," in *Proceedings of ICASSP*, Phoenix, AR, USA, 1999.

[3] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, and W. K. Lo, "A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation," in *Proceedings of COLING*, Geneve, Switzerland, 2004.

[4] N. Bertoldi, R. Cattoni, M. Cettolo, and M. Federico, "The ITC-irst Statistical Machine Translation System for IWSLT-2004," in *Proceedings of IWSLT*, Kyoto, Japan, 2004.

[5] M. Cettolo and M. Federico, "Minimum Error Training of Log-Linear Translation Models," in *Proceedings of IWSLT*, Kyoto, Japan, 2004.

[6] B. H. Tran, F. Seide, and V. Steinbiss, "A Word Graph based N-Best Search in Continuous Speech Recognition," in *Proceedings of ICLSP*, Philadelphia, PA, USA, 1996.

[7] F. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of ACL*, Sapporo, Japan, 2004.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," IBM Research Division, Thomas J. Watson Research Center, Research Report RC22176, 2001.

[9] F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev, "A Smorgasbord of Features for Statistical Machine Translation," in *Proceedings of HLT/NAACL*, Boston, MA, 2004.

[10] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "From Broadcast News to Spontaneous Dialogue Transcription: Portability Issues," in *Proceedings of ICASSP*, Salt Lake City, Utah, USA, 2001.

[11] M. Paul, H. Nakaiwa, and M. Federico, "Towards Innovative Evaluation Methodologies for Speech Translation," in *Working notes of the NTCIR-4 Meeting*, Tokyo, Japan, 2004.

[12] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.