# INVESTIGATING TRANSLATION OF PARLIAMENT SPEECHES

*D. Déchelotte, H. Schwenk, J.-L. Gauvain, O. Galibert, and L. Lamel*

LIMSI-CNRS
BP 133, bat 508, 91436 Orsay cedex, FRANCE
{dechelot,schwenk,gauvain,galibert,lamel@limsi.fr}

## ABSTRACT

This paper reports on recent experiments for speech to text (STT) translation of European Parliamentary speeches. A Spanish speech to English text translation system has been built using data from the TC-STAR European project. The speech recognizer is a state-of-the-art multipass system trained for the Spanish EPPS task and the statistical translation system relies on the IBM-4 model. First, MT results are compared using manual transcriptions and 1-best ASR hypotheses with different word error rates. Then, an $n$-best interface between the ASR and MT components is investigated to improve the STT process. Derivation of the fundamental equation for machine translation suggests that the source language model is not necessary for STT. This was investigated by using weak source language models and by $n$-best rescoring adding the acoustic model score only. A significant loss in the BLEU score was observed suggesting that the source language model is needed given the insufficiencies of the translation model. Adding the source language model score in the $n$-best rescoring process recovers the loss and slightly improves the BLEU score over the 1-best ASR hypothesis. The system achieves a BLEU score of 37.3 with an ASR word error rate of 10% and a BLEU score of 40.5 using the manual transcripts.

## 1. INTRODUCTION

In this paper experiments with an integrated speech to text translation system for European Parliamentary speeches are reported. The task is that of translating European parliamentary speeches from Spanish to English using data from the European TC-STAR (Technology and Corpora for Speech to Speech Translation) integrated project. The speech recognizer is a state-of-the-art multipass system on audio and textual data from the Spanish EPPS task and the statistical translation system relies on the IBM-4 model. The impact of the speech recognizer performance and the quality of the source language model on the translation quality for a real world task is studied.

There has been a recent increase in activity studying the close coupling of speech recognition and translation, both from a theoretical and experimental point of view [8, 16, 13, 17, 12, 7]. In the case of a good translation model, one may argue that the source language model is not necessary for the global task of speech translation. This could not be confirmed in our experiments.

The remainer of this paper is organized as follows. In the following section the task and the available data for developing the speech recognizer and the statistical translation system are described. Sections 3 and 4 summarize the architecture of the speech recognizer and the translation system respectively. The experiments performed in order to get more insight on the coupling of recognition and translation are described in Section 5.

## 2. TASK AND DATA DESCRIPTION

The TC-STAR project,[1] financed by the European Commission within the Sixth Framework Program, is envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (SST). SST technology is a combination of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech synthesis (TTS). The objectives of the project are to significantly reduce the gap between human and machine translation performance.

Within the TC-STAR project, the data consists mainly of texts of the public European Parliament Plenary Sessions (EPPS), held from 1996 to 2004, in English and Spanish, and the corresponding minutes edited by the European Parliament also known as the Final Text Editions [6]. The training portion of the data dates from April 1996 through October 16, 2004. Experimental results reported in this paper make use of the plenary sessions from October 26 and 27, 2004. The final text editions in English and Spanish were aligned at the sentence level. These parallel texts are used to train the statistical machine translation system. Table 1 gives some statistics about the data.

| | Spanish | English |
|---|---|---|
| Sentence Pairs | 1.2M | |
| Total # Words | 31.4M | 30M |
| Vocabulary size | 140k | 94k |
| Singletons | 49k | 34k |

**Table 1**. Statistics of the parallel texts (EPPS final text edition) used to train the statistical machine translation system

Audio recordings are available for a portion of the data from 2004. The sessions from May to October 2004 were transcribed and used to train the speech recognizer (about 40h). A 4h portion was reserved for use as a development test set (see Table 2). The transcriptions of the ASR development test data can also be used to test the translation engine thereby allowing the loss in performance due to ASR errors to be assessed. The resources available to train

THIS WORK WAS PARTIALLY FINANCED BY THE EUROPEAN COMMISSION UNDER THE FP6 INTEGRATED PROJECT TC-STAR.

[1] http://www.tc-star.org/

|  | Train | Dev |
|---|---|---|
| Duration | ≈ 40h | 3.7h |
| Sentences | 27k | 1189 |
| Words | 3.5M | 196k |
| Vocabulary | 27285 | 4018 |

**Table 2**. Statistics of the resources available to build the speech recognizer.

the language models for the speech recognizer are the transcriptions of the training data (3.5M words) and the Spanish final text edition (31.4M words). Within the TC-STAR project, the recognition of the English parliament sessions and their translation to Spanish is also evaluated, but not considered in this paper.

## 3. SPEECH RECOGNITION

The speech recognizer for the Spanish EPPS data uses the same basic modeling and decoding strategy as in the LIMSI English broadcast News system [5]. Details on the adaptation to the EPPS task are given in the following sections.

### 3.1. Models

The acoustic models were trained on about 40 hours of audio training data from the European Parliament speech sessions. The acoustic models are tied-state word position dependent triphones, where the state-tying is obtained using a divisive decision tree clustering. The models used in the first decoding pass include 1700 tied states with 32 Gaussians per state and they cover 1523 triphones. The second pass models cover 5275 triphone contexts and include 5k tied states with 32 Gaussians per state.

A 64k case sensitive vocabulary was used for language modeling. This vocabulary has an OOV rate of 0.6% on the development test data. The language model was built by interpolating three $n$-gram language models trained on the following three data sets: the transcriptions of the acoustic training data (3.5M words), the final text edition of the Parliament sessions for the period 1996-1999, and final text edition for the period 2000-2004. Modified Kneser-Ney smoothing was used as implemented in the SRI LM toolkit [15]. An EM procedure was used to find the interpolation coefficients that minimize the perplexity on the development data. The optimal coefficients are 0.29 for the transcriptions and 0.29 and 0.42 for the two parts of the final text editions.

Since only a limited amount of LM training data was available, a neural network was used to estimate the LM probabilities [1, 14]. The basic idea is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown $n$-grams can be expected. The neural network is used to learn the projection of the words onto the continuous space and to estimate the $n$-gram probabilities. The neural network is trained on the same data as the back-off LM, that is about 35M words (see [14] for more details). The two language models (backoff $n$-gram and neural-net $n$-gram) are interpolated during lattice rescoring.

### 3.2. Decoding

Word recognition is performed in two passes, where each decoding pass generates a word lattice which is expanded with a 4-gram LM. During the first pass initial hypotheses are generated, which are then used for speaker-based acoustic model adaptation. This is done via a one pass (about 1xRT) cross-word trigram decoding with gender-independent sets of position-dependent triphones and a trigram language model. The trigram lattices are rescored with a 4-gram language model. In the second pass, unsupervised acoustic model adaptation of speaker-independent models is performed for each speaker using the CMLLR and MLLR techniques with two regression classes. The segment clusters corresponding to each speaker are obtained automatically using a GMM based clustering procedure [5]. After model adaptation a lattice is generated for each segment using a bigram LM and as for pass 1, the lattices are then rescored with a 4-gram language model.

The lattices of the last decoding pass are rescored by the neural network LM interpolated with the backoff $n$-gram LM. The overall runtime is under than 7xRT. The 4-gram lattices are converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattices edges until a linear graph is obtained, also known as a consensus network. The words with the highest posterior in each confusion set are hypothesized.

|  | Back-off 4-gram | NN 4-gram |
|---|---|---|
| Perplexity | 81.0 | 71.5 |
| Word error, 1-best | 11.1% | 10.2% |
| consensus | 10.6% | 10.0% |

**Table 3**. Performance of the Spanish speech recognizer (automatic segmentation)

Table 3 gives the word error rates of the Spanish speech recognizer with and without consensus decoding, using either a back-off or the neural network language model. Consensus decoding achieves a word error reduction of about 0.5% when used with a back-off LM, but the gain is smaller when the neural network LM is used. Overall the neural network approach achieves a word error reduction of more than 0.6% with respect to the back-off LM. When generating $n$-best lists for close coupling with the translation module, consensus decoding is not used in order to keep the acoustic and language model scores.

The results reported in Table 3 were obtained with a segmentation at long pauses, i.e. not necessarily at sentence boundaries. However current MT evaluation metrics like BLEU, however, suppose that the segmentation corresponds to the sentence structure of the reference translations. Using the segmentation imposed by the reference translations results in a 0.7% loss in the word error rate (The error rate increases from 10.2% to 10.9%).

In the following sections, experiments with various ASR configurations by changing both the acoustic model and the language model are run. The word error rates corresponding to the various conditions are given in Table 4 using the manual MT segmentation.

## 4. STATISTICAL TRANSLATION ENGINE

The so-called IBM-4 [2] model was used in the translation engine. A brief description of this model is given below along with the decoding algorithm.

The translation problem is to find a target sentence $\mathbf{e} = e_1 \dots e_J$ that is a valid translation of a source sentence $\mathbf{f} = f_1 \dots f_I$. Given the fact that the translation models are largely imperfect and are not

| AM/LM | 2-gram | 3-gram | NN 4-gram |
|---|---|---|---|
| Pass 1 models | 16.3% | 14.6% | 13.7% |
| Pass 2 models | 13.5% | 11.8% | 10.9% |

**Table 4**. Word error rates with the first and second pass acoustic models, with 3 different language models (manual segmentation imposed by the reference translation).

guaranteed to produce a grammatical target sentence, the Bayes relation is classically used:

$$\underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e}) \Pr(\mathbf{f}|\mathbf{e}) \tag{1}$$

The search algorithm aims at finding what target sentence $\mathbf{e}$ is most likely to have produced the source sentence $\mathbf{f}$. The translation model $P(\mathbf{f}|\mathbf{e})$ relies on four model components:

1. a fertility model $n(\phi|e)$ that models the number ($\phi_i$) of source words generated by each target word $e_i$;
2. a lexical model of the form $t(f|e)$, which gives the probability that the source word $f$ translates into the target word $e$;
3. a distortion model, that characterizes how words are reordered when translated;
4. and probabilities to model the word insertion of target words that are not aligned to any source words.

The target language model $\Pr(\mathbf{e})$ and the four translation submodels are combined in a log-linear fashion [9], where the combination coefficients are optimized on the development set.

An A* search has been implemented to find the best translation as predicted by the model, when given enough time and memory, i.e. provided pruning did not eliminate it. The decoder manages partial hypotheses, each of which translates a subset of source words into a sequence of target words. Expanding a partial hypothesis consists of covering one extra source position (in random order) and, by doing so, appending one, several or possibly zero target words to its target word sequence.

There are two expansion operators. This first one is called "AddNZFert" and translates exactly one source word by one new target word, preceded by zero or one unfertile target words. In the current implementation, each source word has a maximum of 20 alternative translations, and for each target word $e$, a specific list of 10 target words that both occur frequently before $e$ and are likely to be of fertility 0 is constituted at training time using alignments provided by Giza++. The "AddNZFert" operator thus produces a maximum of $20 * (1 + 10) = 220$ partial hypotheses. The second expansion operator, called "Extend", translates exactly one source word by aligning it to the last target word already produced. The "Extend" operator produces 1 partial hypothesis if applicable, and 0 otherwise.

Partial hypotheses are stored in several queues so as to easily compare hypotheses. For a source sentence of length $J$, $2^J$ queues are created, one per subset of source positions. It is consequently possible to keep those queues relatively small (usual size is 10, a size of 20 gives results marginally different from an infinite stack size) since hypotheses only "compete" with hypotheses that cover the exact same set of source words. This form of pruning is the only one that takes place during decoding. It allows decoding of a 10-word long sentence in two seconds on average, with decoding time almost doubling for each extra source word. To avoid prohibitive decoding times, sentences are automatically split at punctuation signs and at segmental cues (hypothesized punctuation, pauses, breath) given by the speech recognizer, and further split uniformly, if needed.

During decoding, admissible heuristics are used to speed up decoding as well as to accurately prune hypotheses with a high associated probability but whose future expansions are sure to be of low probability. Those heuristics are modeled after [10]. They must be admissible, meaning that they always overestimate the score of the hypothesis future, in order to keep the optimality of the search.

For each source word $f$ at position $i$, an upper bound $P_i$ to the probability of covering $f$ is evaluated as follows. On the one hand, $f$ may be covered by the null target word $e_0$ (which corresponds to no actual target word), with the probability $P_i^0 = t(f|e_0)$. On the other hand, if $f$ is to be translated by a (fertile) target word $e$, the lexical, fertility and distortion submodels should be accounted for. An upper bound to the probability $P_i^+$ of covering $f$ with an actual target word is:

$$P_i^+ = \max_{e,\phi} t(f|e) \sqrt[\phi]{f(\phi|e)} d_{\text{max}} \tag{2}$$

where $d_{\text{max}}$ is constant: it is the maximum value the distortion model can take. Finally, $P_i$ is obtained as the minimum of $P_i^0$ and $P_i^+$:

$$P_i = \min \;\; P_i^0, P_i^+ \tag{3}$$

Decoding uses a 3-gram target language model. Equivalent hypotheses are merged, and only the best scoring one is expanded further.

The BLEU score [11] using two reference translations is the main metric used in this work to evaluate the translation quality. This decoder achieves a BLEU score of 37.6 on the verbatim transcriptions and of 34.6 on the 1-best speech recognition hypothesis with a word error rate of about 10%. This is a relatively small decrease in performance. Different language models were used in the speech recognizer while keeping the same acoustic models. Figure 1 shows BLEU scores against the word error rates of the various systems. As expected, the translation performance decreases with increasing word error rate, although the linear dependence is visually striking.

## 5. INTERACTION BETWEEN TRANSCRIPTION AND TRANSLATION

The overall goal in the European TC-STAR project is to translate speech from one language into spoken text in another language. This can be decomposed into three tasks: recognition of the speech in the source language, translation of the text from the source language to the target language and synthesis of the target language speech. In this work, we concentrate on the close interaction of speech recognition and translation. The simplest way to do speech translation is to first recognize the speech, then produce the most likely hypothesis and to translate it as done in the previous section. In the literature various approaches for closer coupling have been proposed that can be roughly divided into two principles: techniques based on finite-state transducers [16, 3, 4, 7] and techniques based on $n$-best or lattice representations of the transcribed speech signal [13, 17, 12]. For this work an $n$-best interface is used.

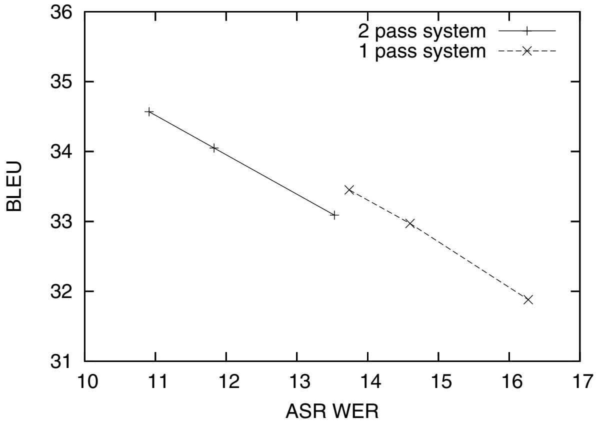

**Fig. 1**. BLEU score with 6 ASR conditions: 2 acoustic models and 3 language models (NN 4-gram, 3-gram, 2-gram) using the MT system and a 1-best ASR-MT interface.

### 5.1. Theoretical background

Let us consider speech translation in the context of Bayes decision theory. We are looking for a target language sentence $\mathbf{e}$ given the acoustic signal $\mathbf{x}$. Following work described by Ney [8], the recognized source text $\mathbf{f}$ is considered as a hidden variable:

$$\begin{aligned}
\mathbf{e}^* &= \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{x}) \\
&= \arg\max_{\mathbf{e}} \Pr(\mathbf{e})\Pr(\mathbf{x}|\mathbf{e}) \\
&= \arg\max_{\mathbf{e}} \quad \Pr(\mathbf{e})\sum_{\mathbf{f}}\Pr(\mathbf{x},\mathbf{f}|\mathbf{e})\Bigg) \\
&= \arg\max_{\mathbf{e}} \quad \Pr(\mathbf{e})\sum_{\mathbf{f}}\Pr(\mathbf{f}|\mathbf{e})\Pr(\mathbf{x}|\mathbf{f},\mathbf{e})\Bigg) \\
&= \arg\max_{\mathbf{e}} \quad \Pr(\mathbf{e})\sum_{\mathbf{f}}\Pr(\mathbf{f}|\mathbf{e})\Pr(\mathbf{x}|\mathbf{f})\Bigg) && (4) \\
&\approx \arg\max_{\mathbf{e}} \quad \Pr(\mathbf{e})\max_{\mathbf{f}}\Pr(\mathbf{f}|\mathbf{e})\Pr(\mathbf{x}|\mathbf{f}) && (5)
\end{aligned}$$

Here the assumption was made that knowing the source $\mathbf{f}$, the target string $\mathbf{e}$ does not help to predict the acoustic observation $\mathbf{x}$, i.e. $\Pr(\mathbf{x}|\mathbf{f},\mathbf{e}) = \Pr(\mathbf{x}|\mathbf{f})$. An important step is to approximate the sum over $\mathbf{f}$ by the maximum. Only in this case we can actually say that there is a recognized source word sequence $\mathbf{f}$. Equation 5 suggests that strictly speaking the source language model $\Pr(\mathbf{f})$ is not necessary for the speech translation task. This motivated the following experimental setup.

### 5.2. Need for a source language model

For each language model and each speech segment a $n$-best list of hypotheses is generated as follows. First, the 1000 most likely different hypotheses are extracted from the lattices generated by the speech recognizer. In practice, many utterances have fewer different hypotheses. Then, hypotheses that are identical for the translation engine are merged, in particular by removing pronunciations variants, filler words and breath noise. Their maximal acoustic score is retained for the resulting hypothesis. At this point, the average sizes of the $n$-best lists varied between 85 and 142 depending on the language models. No attempts were made to optimize their size for translation, the variation in size is a result of the pruning during decoding. All the $n$-best lists are then translated.

Strictly applying equation 5, it should be possible to get the same BLEU score as the 1-best translation by using the score of the acoustic model to rerank the $n$-best translations, and this independently of the source language model used. Hypotheses were obtained using as a cost-function a log-linear combination of the target language model score, the MT and the acoustic model score.

As it can be seen from Figure 2 (by comparing the 2 lower curves ASR-1-best and MT+AM) we were not able to match the 1-best solution by using the acoustic model scores in the log-linear combination, independently of the source LM used to produce the $n$-best lists. It even appears that the gap between the 1-best solution and the rescored $n$-best solution increases when better source language models are used.

### 5.3. Putting back the source language model

From these experimental results it seems clear that the quality of the source language model has an impact on the translation performance. This could be explained as follows. Equation 5 can be rewritten as:

$$\begin{aligned}
\mathbf{e}^* &\approx \arg\max_{\mathbf{e}} \quad \Pr(\mathbf{e})\max_{\mathbf{f}}\Pr(\mathbf{f}|\mathbf{e})\Pr(\mathbf{x}|\mathbf{f}) \\
&= \arg\max_{\mathbf{e}} \quad \max_{\mathbf{f}}\Pr(\mathbf{e},\mathbf{f})\Pr(\mathbf{x}|\mathbf{f}) && (6)
\end{aligned}$$

The translation model $\Pr(\mathbf{f}|\mathbf{e})$ relates a *whole* target sentence to a *whole* source sentence, which implicitly induces a structure on the source sentence. In the case of a simple word translation model, this effect would be rather weak, while stronger relationships can be obtained using alignment templates or by directly estimating the joint probability $\Pr(\mathbf{e},\mathbf{f})$ as suggested in [7]. If the translation model $\Pr(\mathbf{f}|\mathbf{e})$ was perfect, a source language model would probably be not necessary. Note that in text translation the source text $\mathbf{f}$ is given and supposed to be well formed. This is not necessarily the case if we search for all $\mathbf{f}$ maximizing equation 5. Therefore it is wise to add a feature function to the log-linear combination in order to ensure a well formed source sentence before translation. The simplest solution is to add $\log\Pr(\mathbf{f})$ to the feature functions.

This is similar to the experimental setup used in [17, 12] where both ASR and MT generate multiple solutions resulting in $n \times m$-best lists. In these works the primary goal was only to improve upon the 1-best translation, without studying the effect of different source language models. As can be seen from Figure 2, a consistent improvement in the BLEU score with respect to the translation of the 1-best ASR solution was observed (see curve MT+AM+SLM). The gain is 0.8 points in BLEU for the 2-gram source LM and 0.6 points with the neural network LM.

## 6. MT SYSTEM OPTIMIZATION

Since the above experiments were carried out, the MT decoder has been improved by adding support for both independent weighting of the four translation sub-models and producing lattices. The new decoder keeps back-pointers which enables it to dump the search space it explored in the form of a lattice. The lattices are used to tune the coefficients of the log-linear combination. We found that

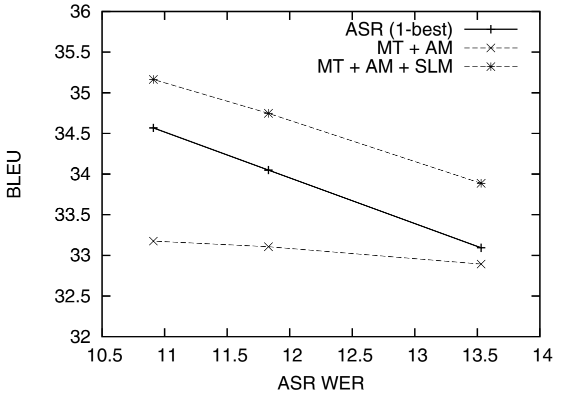

**Fig. 2**. BLEU score when using three different source language models and different log-linear combinations (MT = machine translation scores, AM = acoustic model scores and SLM = source language model score)

the lattice framework is also useful to join sentences that have been split due to their length. Instead of extracting separately the best solution from the lattices corresponding to the different parts of one sentence, all the lattices are sequentially connected and then the global solution is extracted. In our experiments, this gave a gain of 0.3 points in the BLEU score. By tuning the four submodel weights in the log-linear combination, an average 2.5 gain in the BLEU score was observed.

This recently improved version of the MT decoder achieves a BLEU score of 40.5 on the verbatim text and a BLEU score of 37.3 on the 1-best ASR output.

## 7. CONCLUSIONS

This paper described a complete system for speech translation from Spanish to English of public sessions of the European parliament, thus an unconstrained real-world task. The speech recognizer uses the same architecture and decoding procedure as the LIMSI Broadcast News system. The MT system uses the word-based IBM-4 model and implements an A* search with limited pruning. MT results were compared using manual transcriptions and 1-best ASR hypotheses with different word error rates. There is a striking linear dependence of the decrease in translation performance as a function of increasing word error rate.

With a 10% word error rate of the ASR component, the BLEU score is reduced by only about 3% with respect to the score obtained when translating the manual reference transcripts.

Looking at the coupling of ASR and MT, we studied how degraded source language models impact the translation performance and to what extent combining acoustic scores and translation scores can recover the loss in $n$-best rescoring experiments. Combining the acoustic score alone with the translation score, we were unable to rerank the translated $n$-best lists and to match or improve on the translation of the 1-best solution. We explain this by the weakness of the translation model. Using however the complete score of the speech recognizer, i.e. including the acoustic and source language model scores, a consistent improvement in the BLEU score with respect to the 1-best solution was observed. This is consistent with results reported by others [17, 12].

## 8. REFERENCES

[1] Y. Bengio and R. Ducharme. A neural probabilistic language model. In *NIPS*, volume 13, pages 932–938, 2001.

[2] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311, 1993.

[3] F. Casacuberta, D. Llorens, C. Martínex, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pivó, A. Sanchis, E. Vidal, and J. M. Vilar. Speech-to-speech translation based on finite-state transducers. In *ICASSP*, pages 613–616, 2001.

[4] F. Casacuberta and E. Vidal. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225, 2004.

[5] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(2):98–108, 2002.

[6] C. Gollan, M. Bisani, S. Kanthak, R. Schlueter, and H. Ney. Cross domain automatic transcription on the tc-star epps corpus. In *ICASSP*, 2005.

[7] E. Matusov, S. Kanthak, and H. Ney. On the integration of speech recognition and statistical machine translation. In *Eurospeech*, pages 3177–3180, 2005.

[8] H. Ney. Speech translation: Coupling recognition and translation. In *ICASSP*, pages 1149–1152, 1999.

[9] F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*, pages 295–302, University of Pennsylvania, 2002.

[10] F. J. Och, N. Ueffing, and H. Ney. An efficient A* search algorithm for statistical machine translation". In *ACL*, pages 55–62, 2001.

[11] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[12] V. H. Quan, M. Frederico, and M. Catello. Intergrated n-best reranking for spoken language translation. In *Eurospeech*, pages 3181–3184, 2005.

[13] S. Saleem, S.-C. Jou, S. Vogel, and T. Schultz. Using word lattice information for tighter coupling in speech translation systems. In *ICSLP*, pages 422–425, 2004.

[14] H. Schwenk and J.-L. Gauvain. Building continuous space language models for transcribing european languages. In *Eurospeech*, pages 737–740, 2005.

[15] A. Stolcke. SRILM - an extensible language modeling toolkit. In *ICSLP*, pages II: 901–904, 2002.

[16] E. Vidal. Finite-state speech-to-speech translation. In *ICASSP*, pages 111–114, 1997.

[17] 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 *Cooling*, 2004.