SPEECH TRANSLATION BASED ON AUTOMATICALLY TRAINABLE FINITE-STATE MODELS*

J. C. Amengual[1] *J. M. Benedí*[2] *K. Beulen*[3] *F. Casacuberta*[2] *A. Castaño*[1]
A. Castellanos[1] *V. M. Jiménez*[2] *D. Llorens*[2] *A. Marzal*[1] *H. Ney*[3]
F. Prat[1] *E. Vidal*[2] *J. M. Vilar*[1]

(1) Unidad Predepartamental de Informática
Campus Penyeta Roja, Universitat Jaume I
E-12071 Castelló (Spain)

(2) Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
E-46071 Valencia (Spain)

(3) Lerhstuhl für Informatik VI
RWTH Aachen, University of Technology
D-52056 Aachen (Germany)

e-mail: evidal@iti.upv.es

ABSTRACT

This paper extends previous work exploring the use of Subsequential Transducers to perform speech-input translation in limited-domain tasks. This is done following an integrated approach in which a Subsequential Transducer replaces the input-language model of a conventional speech recognition system, and is used both as language and translation model. This way, the search for the recognised sentence also produces the corresponding translation. A corpus-based approach is adopted in order to build the required models from training data. Experimental results are presented for the translation task considered in the EUTRANS project: one in the hotel domain with more than 500 words per language and language perplexities near to 10.

1. INTRODUCTION

A Subsequential Transducer (SST) is a deterministic finite-state machine that accepts sentences from a given input language and produces associated sentences of an output language [1]. Speech-input language translation by means of SSTs can be done following an integrated approach in which the input-language model of a conventional speech recognition system [2] is replaced by a SST, which is used both as language and translation model. This way, the search for the recognised sentence also produces the corresponding translation [3].

A Corpus-Based (CB) approach can be adopted in order to build the required SSTs from training data. Different CB approaches to Machine Translation (MT) have emerged in the last few years [4–8] as an alternative to more traditional, knowledge-based systems. An important advantage of SSTs is that they can be automatically learnt from a representative set of training examples (consisting in pairs of sentences that are translations of each other). If the corresponding function is total, the algorithm known as OSTIA [9] can be used. If the function is partial and a finite state model restricting its domain (input language)—and possibly other for its range (output language)—is provided, an extension of OSTIA known as OSTIA-DR [10] can be employed. The incorporation of input and output language restrictions while learning the SSTs becomes especially important when the SSTs are going to be used both as language and translation models [3]. In particular, the domain and range models used with OSTIA-DR can be stochastic n-testable automata, or equivalently n-grams [11].

The use of automatically learnt SSTs for limited-domain translation tasks showed useful performance in preliminary experiments with an academic task involving descriptions of visual scenes [3, 12]. In order to deal with more complex tasks, some extensions to our basic speech-to-speech translation system have been implemented. In particular, acoustic modelling has been improved by using continous-density (instead of discrete) Hidden Markov Models (HMMs) and the grouping of some words and word sequences into categories has been introduced in order to enhance the SST learning process. In this work, the translation system built in that way is tested upon a speech-input translation task in the hotel domain.

2. WORD CATEGORIES

The approach presented in [13] for the integration of categories with SSTs proved that their use is very effective in reducing the size of the final transducers and the amount of training data needed. However, this approach was not easily integrable in a speech recognition system and did not consider the possibility of having categories including units larger than a word (like numbers).

The approach presented here circunvents these problems creating a single SST that comtains all the information

*Work partially funded by the Spanish C.I.C.Y.T. (project TIC95-0984-C02) and by the European Union (ESPRIT project no. 20268).

necessary for the translation, including the basic transducers for the categories. The main steps are:

- A set of categories is identified.
- The training corpus is categorized accordingly and used for training an initial model.
- For each category, a simple SST is built.
- The arcs in the initial model corresponding to the different categories are expanded using these simple SSTs.

Seven categories were used: masculine names, femenine names, surnames, dates, hours, room numbers and general numbers. These categories follow quite simple translation rules and the amount of linguistic knowledge introduced by them is very low.

The categorization of the corpus was done by replacing the words and word sequences by adequate labels and using this labelled corpus as training material. This implies that the corresponding SST had arcs labelled by category names instead of vocabulary words. In principle, it is possible to expand the transducers for the categories in the corresponding arcs so that the final translation is obtained directly, but this expansion can be complex and generate too large models. The approach we followed was to keep the labels in the translation together with information on their substitution. For example, let us suppose the Spanish sentence:

> *Por favor, dénos las llaves de la doscientos veintidós.*

Then, the expanded SST will produce:

> *Please give us the keys to room number $ROOM*
> *$ROOM=[two two two].*

A simple postprocess is used to obtain the final translation. A more detailed explanation of the process can be found in [14].

3. THE TRAVELER TASK

The *Traveler task* [15] was defined within the EUTRANS project [8] in order to be a more realistic test for our CB MT techniques than that we previously employed [12]. This task is aimed at both being realistic and allowing fast, cost-effective, automatic data generation. The general framework adopted for the task is that of a traveler (tourist) visiting a foreign country whose language he/she does not speak. This framework includes a great variety of different translation scenarios, and thus results appropriate for progressive experimentation with increasing level of complexity. In a first phase, the scenario has been limited to some human-to-human communication situations in the reception of a hotel:

- Asking for rooms, wake-up calls, keys, the bill, a taxi and moving the luggage.
- Asking about rooms (availability, features, price).
- Having a look at rooms, complaining about and changing them.
- Notifying a previous reservation.
- Notifying the departure.
- Asking and complaining about the bill.
- Signing the registration form.
- Other common expressions.

The *Traveler task* text-to-text corpora are sets of pairs each of which consists in a sentence in the input language and the corresponding translation in the output language. These sets were automatically built, since automatic generation allows the obtainment of paired samples in a large enough quantity for making CB MT experiments possible [16]. Moreover, the complexity of the task can be controlled. On that score, the existing MT corpora are either too restricted (such as those related to the *Miniature Language Acquisition task* originally introduced in [17] and adequately reformulated later as a MT task [12]) or unrestricted data (such as the Hansard corpora [4]).

In order to generate Spanish-to-English text-to-text data for the *Traveler task*, a set of Stochastic, Syntax-directed Translation Schemata (SSTSs) [18] was developed. These SSTSs were then used to automatically generate a huge corpus of pairs of sentences through a data generation tool, specially developed for the EUTRANS project. This software allows the use of several syntactic extensions to SSTS specifications in order to express optional rules, permutation of phrases, concordance (of gender, number and case), etc.

Some examples of the resulting Spanish-to-English translations are shown in Figure 1, and Table 1 summarizes the main features of this corpus. For each language, the test set perplexity has been computed by training a trigram model (with simple flat smoothing) using a set of 20,000 random sentences and computing the probabilities yielded by this model for a set of 10,000 independent random sentences. The lower perplexity of the output language derives from a design decision: multiple variants of the input sentences were introduced to account for different ways of expressing the same idea, but they were given the same translation.

Finally, a speech corpus for the task was built. A total 436 Spanish sentences were selected from the text corpus. They were divided into eleven sets:

Spanish:	*Quisiéramos reservar dos habitaciones para un día a nombre de Federico Mestre, por favor.*
English:	*We want to book two rooms for a day for Federico Mestre, please.*
Spanish:	*Por favor, dénos las llaves de la doscientos veintidós.*
English:	*Please give us the keys to room number two two two.*
Spanish:	*¿Cuánto cuesta por día una habitación doble con pensión completa?*
English:	*How much does a double room with full board cost per day?*

Figure 1: Some examples of Spanish-to-English translations in the *Traveler task*.

Table 1: Main features of the Spanish-to-English text corpus.

	Spanish	English
Vocabulary size	689	514
Average sentence length	9.5	9.8
Test-set perplexity	13.8	7.0
Sentence pairs (Different)	500,000 (171,481)	

- One common set consisting of 16 sentences.
- Ten sets of 42 sentences.

Each of the twenty speakers (ten male and ten female) participating in the acquisition of this corpus, pronounced the common set and two out of the other ten, totalling 2,000 sentences, 15,360 words and about 90,000 phones. The sampling frequency was 16 kHz.

From this speech corpus, two subcorpora were extracted:

- Training and adaptation (*TravTR*): 16 speakers (eight male and eight female), 268 sentences, 1,264 utterances (aprox. 11,000 words or 56,000 phones).
- Speaker independent test (*TravSI*): 4 speakers (two male and two female, not in *TravTR*), 84 sentences (not in *TravTR*), 336 utterances (aprox. 3,000 words or 15,000 phones).

4. EXPERIMENTAL RESULTS

In the experimental results presented here, each word of the Spanish vocabulary has been modelled as a simple concatenation of phones (from a set of 31 that includes stressed and unstressed vowels plus two types of silence), which in Spanish can be derived from standard phonetic rules. The acoustic modelling of these phones has been carried out using context-independent continuous-density HMMs [19–21] whose parameters had been estimated using the union of two corpora: the 1,264 utterances of the *TravTR* subcorpus, together with a small set of 1,530 utterances (by 9 speakers, 4 male and 5 female) from a different task that was designed in order to have a quasi phonetically-balanced corpus. This speech material was processed to obtain, each 10 msecs, 10 cepstral coefficients of a Mel-filter bank plus the energy and the corresponding first and second derivatives. A training set of 168,629 different Spanish-to-English (text) sentence pairs from the *Traveler task* was used to automatically learn 3-gram models for the input and output languages. A SST was then learnt with OSTIA-DR using these input and output 3-grams and the same training pairs.

Table 2: Spanish-to-English speech-input recognition and translation results.

Number of Gaussians	AMBF	Recognition WER	Translation WER	RTF
1,663	300	2.7 %	2.3 %	5.9
	150	6.6 %	6.4 %	2.2
5,590	300	2.2 %	1.9 %	11.3
	150	6.7 %	6.3 %	5.6

The system was afterwards used to recognise and translate into English the 336 Spanish utterances of the *TravSI* subcorpus*. The search was performed using the Viterbi algorithm with a beam search at two levels: independent beam widths were used in the states of the SST (Language Model Beam Width—LMBW) and in the states of the HMMs (Acoustic Model Beam Width—AMBW). The scores obtained in the two levels were linearly combined in the log scale using empirically obtained factors. The LMBW was empirically fixed at 300. Table 2 presents the recognition and translation Word Error Rate (WER) and Real Time Factor (RTF) achieved on a HP-9735 workstation, for different number of Gaussian distributions and different AMBFs.

Thanks to the lower perplexity of the output language and to the integrated approach here adopted (in contrast with traditional decoupled approaches in which the output of a speech recogniser is submitted to a different translation module) the translation WER is slightly lower than the recognition WER. Also, there is a clear tradeoff between computing time and accuracy. For the models with 1,663 Gaussian distributions and tight beam-search thresholds, the recognition and translation computing-time is only 2.2 times real time without resorting to any type of specialised hardware or signal processing device. This configuration provides an acceptable behaviour for on-line operation. For off-line operation, a different configuration can provide improved performance at the cost of increasing the RTF.

*It is worth noting that the speakers and sentences used in training the translation and acoustic models are disjoint from those in *TravSI*.

5. CONCLUSION

The use of SSTs allows the integration of acoustic and translation models in the building of speech-to-speech translation systems for medium sized tasks. Good recognition rates are achieved by using continuous HMMs and more compact SSTs are obtained by using categories, for words and word sequences, in the training corpora.

Future directions include reducing the number of sentences necessary for training the translation models in order to cope with spontaneous instead of synthetic sentences. For this, new approaches are being explored, like reordering the words in the translations, the use of new inference algorithms, and automatic categorization.

REFERENCES

[1] Berstel, J. 1979. *Transductions and Context-Free Languages*. Stuttgart: Teubner.

[2] Ney, H. 1995. "Search Strategies for Large-Vocabulary Continuous-Speech Recognition". In A. J. Rubio and J. M. López (eds.), *Speech Recognition and Coding, New Advances and Trends*, pp. 210–225. NATO Advanced Study Institute. Berlin: Springer-Verlag.

[3] Jiménez, V. M., Castellanos, A., and Vidal, E. 1995. "Some Results with a Trainable Speech Translation and Understanding System". Proceedings of the ICASSP-95, Detroit (USA), pp. 113–116.

[4] Brown, P. F. *et al*. 1990. "A Statistical Approach to Machine Translation". *Computational Linguistics* **16**(2): 79–85.

[5] Vidal, E., Pieraccini, R., and Levin, E. 1993. "Learning Associations between Grammars: A New Approach to Natural Language Understanding". Proceedings of the EUROSPEECH-93, Berlin (Germany), pp. 1187–1190.

[6] Castaño, M. A., and Casacuberta, F. 1997. "A Connectionist Approach to Machine Translation". In these proceedings.

[7] Tillmann, C., Vogel, S., Ney, H., and Zubiaga, A. 1997. "A DP-Based Search using Monotone Alignments in Statistical Translation". To appear in Proceedings of the ACL-EACL'97 Joint Conference, Madrid (Spain).

[8] Amengual, J. C. *et al*. 1996. "EuTrans: Example-Based Understanding and Translation Systems: First-Phase Project Overview". Technical Report D4, Part I, EuTrans (ESPRIT project no. 20268). (Restricted.)

[9] Oncina, J., García, P., and Vidal, E. 1993. "Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks". *IEEE Transactions on PAMI* **15**(5): 448–458.

[10] Oncina, J., and Varó, M. A. 1996. "Using Domain Information During the Learning of a Subsequential Transducer". In L. Miclet and C. De La Higuera (eds.), *Grammatical Inference: Learning Syntax from Sentences*, pp. 301–312. Lecture Notes in Artificial Intelligence 1147. Berlin: Springer-Verlag.

[11] Vidal, E., Casacuberta, F., and García, P. 1995. "Grammatical inference and automatic speech recognition". In A. J. Rubio and J. M. López (eds.), *Speech Recognition and Coding, New Advances and Trends*, pp. 174–191. NATO Advanced Study Institute. Berlin: Springer-Verlag.

[12] Castellanos, A., Galiano, I., and Vidal, E. 1994. "Application of OSTIA to Machine Translation Tasks". In R. C. Carrasco and J. Oncina (eds.), *Grammatical Inference and Applications*, pp. 93–105. Lecture Notes in Artificial Intelligence 862. Berlin: Springer-Verlag.

[13] Vilar, J. M., Marzal, A., and Vidal, E. 1995. "Learning Language Translation in Limited Domains using Finite-State Models: Some Extensions and Improvements". Proceedings of the EUROSPEECH-95, Madrid (Spain), pp. 1231–1234.

[14] Amengual, J. C. *et al*. 1997. "Using Categories in the EuTrans System". To appear in Proceedings of the EACL Spoken Language Translation Workshop, Madrid (Spain).

[15] Amengual, J. C. *et al*. 1996. "Definition of a Machine Translation Task and Generation of Corpora". Technical Report D1, EuTrans (ESPRIT project no. 20268). (Restricted.)

[16] Vidal, E. 1997. "Finite-State Speech-to-Speech Translation". Proceedings of the ICASSP-97, Munich (Germany), pp. 111–114.

[17] Feldman, J. A., Lakoff, G., Stolcke, A., and Weber, S. H. 1990. "Miniature Language Acquisition: A Touchstone for Cognitive Science". Technical Report TR-90-009, International Computer Science Institute, Berkeley (USA).

[18] Gonzalez, R. C., and Thomason, M. G.. 1978. *Syntactic Pattern Recognition: An Introduction*. Reading: Addison-Wesley.

[19] Haeb-Umbach, R., and Ney, H. 1994. "Improvements in Time-Synchronous Beam-Search for 10000-Word Continuous Speech Recognition". *IEEE Transactions on SAP* **2**(3): 353–356.

[20] Beulen, K., Welling, L., and Ney, H. 1995. "Experiments with Linear Feature Extraction in Speech Recognition". Proceedings of the EUROSPEECH-95, Madrid (Spain), pp. 1415–1418.

[21] Welling, L., Ney, H, Eiden, A., and Forbrig, Ch. 1995. "Connected Digit Recognition using Statistical Template Matching". Proceedings of the EUROSPEECH-95, Madrid (Spain), pp. 1483–1486.