FINITE-STATE SPEECH-TO-SPEECH TRANSLATION*

Enrique Vidal
Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, 46020 Valencia, SPAIN

ABSTRACT

A fully integrated approach to Speech-Input Language Translation in limited-domain applications is presented. The mapping from the input to the output language is modeled in terms of a *finite state* translation model which is learned from examples of input-output sentences of the task considered. This model is tightly *integrated* with standard acoustic-phonetic models of the input language and the resulting global model directly supplies, through Viterbi search, an optimal output-language sentence for each input-language utterance. Several extensions to this framework, recently developed to cope with the increasing difficulty of translation tasks, are reviewed. Finally, results for a task in the framework of hotel front-desk communication, with a vocabulary of about 700 words, are reported.

1. INTRODUCTION

Language Translation (LT) has been among the main objectives of the work carried out by the Linguistic Research community over the last few decades, and a number of *Text-input* LT (TLT) systems have been developed for practical use. On the other hand, current *Speech Recognition* technology is already sufficiently developed to be used in many speech recognition applications. Hence, the most straightforward approach to achieve *speech-input* LT (SLT) is to serially couple an input speech recognition front-end followed by a conventional TLT program. However, these conventional TLT systems are by no means perfect and a great deal of false responses and lack of coverage occurs in real applications. Clearly, if such inaccuracies are (serially) combined with the also imperfect behavior of our speech-input front-ends, only poor overall performance can be expected.

These considerations suggest that SLT needs a much more *integrated approach*. In attempting to develop such an approach, we try to formulate the problem under a framework that is closer to the standard assumptions under which our successful speech recognition systems are currently developed. This means i) to devise *simple* and easily understandable models for LT, ii) to formulate (S)LT as some kind of *optimal search* through an adequate structure based on these models, and iii) to develop techniques to actually *learn* the LT models from training data of each considered task. All these requirements can be easily met through the use of *Finite-State Translation Models*.

The capabilities of Finite-State Models (FSM) have been the object of much debate in the past few years. On the one hand, in the Natural Language (NL) community, FSMs have often been ruled out for many NL processing applications, including MT, even in limited domains. On the other hand, in Speech Recognition, the use of N-Gram models, which are just among the simplest types of FSMs [17, 8, 18], is firmly established nowadays as a state of the art technique for Language Modeling, even for open-domain applications. Recently, many NL and Computational Linguistic researchers are (re-)considering the interesting features of FSMs for their use in NL processing applications [10].

Simple as they are, FSMs generally need to be huge in order to be useful approximations to complex languages. For instance, an adequate 3–Gram Language Model for the language of the Wall Street Journal is a FSM that may have as many as 20 million edges [15]. Obviously, there is no point in trying to manually build such models on the base of a priori knowledge about the language to be modeled: the success lies in the possibility of automatically learning them from large enough sets of training data [8, 15]. This is also the case for the finite-state translation models used in the work presented in this paper [11, 16, 17].

2. SUBSEQUENTIAL TRANSDUCERS

The difficulty of a translation task depends on many factors. One of the most important is the *"asynchrony"*, or distance at which words of the output sentences tend to appear with respect to their corresponding input-language words. A conventional hierarchy that emphasizes this view, along with some examples of translation tasks (possibly) belonging to each to each level of the hierarchy are shown in Figure 1 [16].

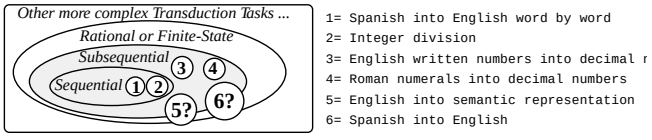

Figure 1. Some interesting classes of Formal Transduction and examples of real tasks (possibly) belonging to these classes.

Particularly interesting for our purposes is the class known as *Subsequential Transduction* [2]. In the transductions of this class, output symbols or substrings are generated only after having seen enough input symbols to guarantee a correct output. The amount of symbols to wait for may be variable and context-dependent and it can even be necessary to append some additional symbol or substring to

* Work partially supported by the Spanish CICYT under grant TIC95–0984–C02–01 and by the European Commission under contract 20268 (ESPRIT project EUTRANS). Many people have contributed to different parts of this work: Juan-Carlos Amengual, José-Miguel Benedí, Francisco Casacuberta, Asunción Castaño, Antonio Castellanos, Víctor M. Jiménez, David Llorens, Andrés Marzal, Federico Prat, Héctor Rulot and Juan Miguel Vilar from the *Instituto Tecnológico de Informática*, Valencia (SPAIN); Cristina Delogu and Andrea Di Carlo from the *Fondazione Ugo Bordoni*, Roma (ITALY), Hermann Ney and Stephan Vogel from the Aachen *Rheinisch-Westfälische Technische Hochschule Lehrstuhl für Informatik VI* (GERMANY). Thanks are also due to CNET (France Telecom) and, in particular, to Christel Sorin and her team for making available the English and German Text–To–Speech synthesizers used in the speech–to–speech prototypes developed in this work.

the output string which can only be determined after having detected the end of the input string.

It should be noted that many translation tasks that may appear much more complex, are inherently of this subsequential nature. For instance, we can always translate natural English sentences into correct Spanish by successively outputting Spanish words that can be determined from a finite (often short) sequence of previously seen English words. In other words, we do not need to wait for a whole discourse to end before starting the translation.

A *Subsequential Transducer* (SST) is a deterministic network having a *finite set of states* and a set of *edges or transitions* linking these states. Each transition is labeled with a (non-empty) *input symbol* and a (possibly empty) string of *output symbols*. Each (final) state is also labeled with a (possibly empty) string of output symbols. The translation of an input string, x, is produced by concatenating the output strings of the edges used to parse the successive symbols of x and finally appending the output string associated to the last state reached through the parsing [2].

Apart from their adequateness for many language processing applications [10], most attractive is the fact that SSTs are *learnable* from training input-output sentence pairs, using a very efficient algorithm called *Onward Subsequential Transducer Inference Algorithm (*OSTIA*)* [11].

SSTs and OSTIA have been successfully used in a variety of applications, including Language Understanding [4, 14, 16], and LT as will be discussed below.

3. AN EXPERIMENTAL TASK: MLA

Most of the early work on the application of SSTs and OSTIA to LT was carried out by testing the different concepts and techniques on a simple and flexible Language Learning task originally introduced by Feldman and his collaborators. The task is called "Miniature Language Acquisition" (MLA) and involves description and manipulation of simple visual scenes [6]. This task, which has proved very useful for experimentation purposes, was adequately reformulated as a LT task. The original formulation involved fairly simple syntax and small lexicon (about 30 words), but it was extended, as required, in order to study the impact of increasing degree of input-output *asynchrony, vocabulary size*, etc. [5, 19]. A large corpus of Spanish-English paired sentences of this task was generated semi-automatically [5], following the syntax directions dictated by the original formulation of the task [6], as well as by the adopted extensions. Examples of MLA sentences are shown in Figure 2.

Spanish:	un cuadrado mediano y claro y un círculo tocan a un círculo claro y un cuadrado mediano
English:	a medium light square and a circle touch a light circle and a medium square
Spanish:	se elimina el círculo grande que esta encima del cuadrado y del triángulo mediano
English:	the large circle which is above the square and the medium triangle is removed

Figure 2. Spanish-English sentences from the MLA task.

4. MLA TRANSLATION EXPERIMENTS

4.1. Basic Text-Input Results

A first series of experiments[1] were carried out with the simplest versions of the MLA translation task using the basic version of OSTIA [11]. The results of these experiments were encouraging. Performance of the learned transducers gradually improved with the amount of training data. In the case of Spanish-to-English, translation accuracy higher than 99% was obtained with transducers learnt from 16,000 training pairs, and almost perfect results were achieved by training with 50,000 pairs. These transducers were very small, typically less than 20 states and 200 edges, and learning time was less than 100 seg. on a HP9715/35 computer. English-to-Spanish proved somewhat more difficult, requiring more than 50,000 pairs to reach 99% accuracy [5].

[1] Only Spanish-English experiments are reviewed here. German output was also considered with similar results [5, 12, 9].

4.2. Speech-Input Experiments: Learning with Input-Output Language Model Constraints

The learning strategies followed by OSTIA try to generalize the training pairs as much as possible. As discussed in the last section, this often leads to very compact transducers that accurately translate *correct* input text. However, this compactness often entails excessive *over-generalization* of the input and output languages, allowing nearly meaningless input sentences to be accepted, and translated into even more meaningless output! While this is not actually a problem for perfectly *correct text* input, it leads to dramatic failures when dealing with not exactly correct text or (even "correct") *speech* input.

A possible way to overcome this problem is to limit generalization by imposing adequate Language Model (LM) constraints: the learned SSTs should *not* accept input sentences or produce output sentences which are not consistent with given LMs of the input and output languages. These LMs are also known as *Domain* and *Range* models [13]. Learning with Domain and/or Range constraints can be carried out with a version of OSTIA called OSTIA-DR [12, 13]. This version was used in [9] in a series of Spanish-English MLA *speech-input* translation experiments.

To deal with acoustic input, very simple, 3-state acoustic-phonetic discrete Hidden Markov Models, trained with Spanish speech from other applications, were used to represent (expand) the input words associated to each edge of the learned transducers.

Transducer learning was based on $50,000$ randomly selected Spanish–English training pairs of the task. From these pairs, 4–Gram Input and Output LMs were obtained from the input and output sentences, respectively, and then OSTIA(-DR) was used to learn *four* different translation models: the first one was learnt with the original OSTIA (no input or output LMs); in the second case, for the sake of comparison, a *"decoupled"* (as opposed to *integrated*) architecture was tried in which a conventional speech recognition front-end, having the input 4–Gram as its LM, was used to recognize the input utterances and the recognized sentences were submitted to translation by the transducer obtained by OSTIA; the third and fourth models were fully *integrated* models learnt with OSTIA-DR, constrained by either the input 4–Gram or both input and output 4–Grams, respectively. These integrated SSTs were not large; typically less than 100 states and 400 edges.

Testing was carried out using 100 random Spanish sentences (none of them used as training), uttered by 4 different speakers (400 utterances in total). Although integrated transducers directly provide only output-language translations, the corresponding input-language sentences are easily obtained as a by-product. This allows us to measure both the "recognition" and translation Word Error Rates for these models. Table 1 shows the results

Table 1. Speech-Input Recognition and Translation Word Error Rates (in %) for different usages of Input and Output 4–Gram Language Models.

Language Model Usage	Recog.	Trans.
No LMs (only the basic SST)	98.0	96.5
Decoupled (Input LM front-end)	4.8	15.2
Integrated: Input LM Only	3.5	3.7
Integrated: Input/Output LMs	2.6	2.8

As expected, the original transducer was completely useless for speech-input operation. The speech recognition

front-end with the 4–Gram input LM led to better results but, even with a relatively high *recognition* accuracy (4.8% WER), large *translation* errors were produced (15.2% WER). This lack of *translation robustness* improved dramatically when integrated models were used, and more so as more syntactic constraints were taken into account.

It is worth noting that the *speech recognition* results obtained for the integrated transducers are significantly better than those achieved by the (*decoupled*) front-end using the very same input-language 4–Gram. This means that the integrated transducers offer better "implicit" modeling of the input language than the input 4–Gram itself.

4.3. Reducing the demand for training data

The amount of training data required by OSTIA(-DR)–learning is directly related with the size of the vocabularies and the amount of input-output *asynchrony* of the translation task considered. This is due to the need of "delaying" the output until enough input has been seen. In the worst case, the number of states required by a SST to achieve this delaying mechanism can grow as much as $O(n^k)$, where n is the number of (functionally equivalent) words and k the length of the delay.

Techniques to reduce the impact of k were studied in [20, 21]. The proposed methods rely on *reordering* the words of the (training) output sentences on the base of *partial alignments* obtained by statistical translation methods [3]. Obviously, adequate mechanisms are provided to recover the correct word order for the translation of new test input sentences [21]. Using these techniques under the same conditions as those of the experiments reported in Table 1 for integrated models, the results shown in Table 2 were obtained. Only 4,000 training pairs were required to achieve similar results as the direct approach with at least four times more training data [20].

Table 2. Impact of Word Reordering on model size (Edges) and Speech–Input translation Word Error Rate (WER in %) for increasing size of the training–set supplied to OSTIA-DR with input/output 4-gram constraints.

	Direct		Reordered	
Train.Pairs	Edges	WER	Edges	WER
1,000	2,023	45.5	1,338	17.5
2,000	3,353	38.7	979	7.3
4,000	4,051	28.2	440	3.2
8,000	719	4.8	344	3.0
16,000	363	3.2	183	3.3

On the other hand, techniques to cut down the impact of vocabulary size were studied in [19]. The basic idea was to substitute words or groups of words by labels representing their syntactic (or semantic) *category* within a limited rank of options. Learning was thus carried out with the categorized sentences, which involved a (much) smaller effective vocabulary. Obviously, categorization has to be done for input/output *paired clusters*; therefore adequate techniques are needed to represent the actual identity of input *and* output words in the clusters and to recover this identity when parsing test input sentences. Text–Input experiments using these techniques were presented in [19] for extended versions of MLA. A summary of the results is shown in Table 3. While the direct approach degrades rapidly with increasing vocabulary sizes, categorization keeps the accuracy essentially unchanged.

5. A MORE COMPLEX AND PRACTICAL APPLICATION: THE "TRAVELER TASK"

After the basic studies carried out with the experimental MLA task, a more ambitious and practically motivated task has recently been considered [1]. The general domain is that of a traveler (tourist) visiting a foreign country. It encompasses a variety of different translation scenarios which range from limited-domain applications to unrestricted natural language. This allows for progressive experimentation with increasing level of complexity. For the results reported below, the domain has been limited to human-to-human communication situations in the front-desk of a hotel.

Table 3. Impact of Categorization on Text-Input translation Sentence Error Rate (in %) for two training-set sizes and increasing vocabulary sizes.

Inp/Out	8,000 Tr.Pairs		32,000 Tr.Pairs	
Voc.Sizes	Direct	Categ.	Direct	Categ.
37/28	3.1	0.9	0.5	0.2
50/38	42.1	1.5	5.7	0.3
63/48	62.5	3.0	26.5	0.6
363/248	91.3	3.4	98.0	0.7

In order to define more precisely the chosen task, several traveler-oriented booklets were collected and those pairs of sentences fitting the above scenario were selected. This provided a (small) "seed corpus" from which a large set of sentence pairs was generated in a semi-automatic way [1]. Data was generated for the following language pairs: Spanish-English, Spanish-German and Spanish-Italian. Table 4 shows some features of the first of these corpora[2], along with examples of paired sentences. Note that output-language perplexity is lower than that of the input language. This reflects a real feature of the corpus: to be useful in practice, the translated sentences need not to exhibit the same linguistic variability as the corresponding input sentences. Obviously, the translation system should be prepared to accept all the relevant input variability, but should only produce correct, simple and concise output.

Table 4. Features of the Traveler Task Spanish-English Corpus and some examples of input-output sentences.

Different sentence pairs in the corpus	171,481
Input/output vocabulary sizes	689 / 514
Average input/output lengths	9.5 / 9.8
Input/output test-set perplexities	13.8 / 7.0

Spanish:	Reservé una habitación individual y tranquila con televisión hasta pasado mañana.
English:	I booked a quiet, single room with a tv. until the day after tomorrow.
Spanish:	Por favor, prepárenos nuestra cuenta de la habitación dos veintidós.
English:	Could you prepare our bill for room number two two two for us, please?

6. TRAVELER TASK EXPERIMENTS

6.1. Text-Input Experiments

From the corpus discussed in the previous section, increasing amounts of randomly selected training pairs[3] were used to study the learning convergence of OSTIA-DR with input and output 3-Gram LM constraints. Testing was carried out on 2,730 different input sentences which were not seen in training. Two types of experiments were carried out. In the first one, OSTIA-DR was directly used. In the second, OSTIA-DR was assisted by categorization techniques similar to those discussed in Section 4.3. In this case, seven categories were adopted, including *room numbers, dates, times-of-day, names, surnames*, etc. The results are shown in Table 5. Useful accuracy was obtained starting with transduc-

[2]Only Spanish-English experiments will be reported here; similar behavior was observed for the other language pairs [1]. Perplexity figures correspond to a standard (flat smoothed) trigram model trained from a set of 20,000 randomly selected sentences and tested with 10,000 independent sentences.

[3]Many (simple) training sentences like "Good morning", "Thank you" etc, appeared many times in the corpus. The repetitions were removed for OSTIA-DR learning but not for probability estimation.

ers learnt with about 30,000 different categorized training pairs. The sizes of these transducers were quite affordable: less than 4,300 states and 40,000 edges.

Table 5. Traveler Task Text-Input translation WER (in %) using input and output 3–Gram LM constraints in OSTIA–DR learning, with and without lexical categorization.

Training Pairs		Translation WER in %	
Different	Categorized	Direct	Categ.
12,218	9,981	54.9	22.5
21,664	16,207	47.9	13.7
38,438	25,665	38.4	7.7
67,492	39,747	26.0	3.7
119,048	60,401	17.4	1.4
168,629	77,499	13.3	0.7

6.2. Speech-Input Experiments

The best Spanish-English transducer obtained in the previous text-input experiments was used. The Spanish words associated to the edges of the transducer were modeled as a simple concatenation of phonetic elements, from a set of 31 (context-independent) units, including stressed and unstressed vowels plus two types of silence. These units were modeled by context-independent continuous-density Hidden Markov Models [7], whose parameters were estimated using Spanish speech data from other applications (10,700 words by 10 speakers), along with Spanish sentences of the Traveler Task (11,000 words by 16 speakers). The test–set consisted of 84 Traveler Task sentences, uttered by 4 speakers (336 utterances –3,000 words approx.). None of these test sentences or speakers were used in OSTIA-DR or HMM training. Table 6 summarizes the results.

Table 6. Traveler Task speech–input results: recognition and translation Word Error Rate (WER, in %), and computing Real Time Factor (RTF) on a HP-9735 workstation, for different number of Gaussian distributions and beam width constants.

Number of Gaussians	Beam Width	Comp. RTF	Recog. WER	Trans. WER
1663	200	3.2	4.5	4.6
1663	300	5.9	2.7	2.3
5590	300	11.3	2.2	1.9

Using tight beam-search thresholds, an adequate overall performance is obtained with conventional UNIX workstations, without resorting to any type of specialized hardware or signal processing device. As assessed by on-line tests with an implemented prototype [1] (which produced speech output using a state-of-the-art text-to-speech synthesizer made available by CNET – France Telecom), this provides quite acceptable behavior for practical use.

As expected, a high degree of combined recognition/translation *robustness* is achieved. In fact, translation WER tends to be slightly lower than recognition WER. This is consistent with the relative perplexities of the input and output languages reported above Table 4): the tight integration of acoustic, syntactic and translation models allows taking advantage of the lower perplexity of the output language to actually improve the overall accuracy! Note that this behavior is in contrast with the degradation of results that often plague current translation approaches which base their operation on a loosely coupled "first-recognition-then-translation" paradigm.

7. CONCLUDING REMARKS

Finite-State Formal Transduction techniques have been proposed for language translation in limited domains, and novel techniques for *learning* the required transducers have been reviewed. Results show great potential for developing simple and effective systems for *limited-domain speech-to-speech language translation applications.*

REFERENCES

[1] J.C.Amengual et al. "Example-Based Understanding and Translation Systems (EuTrans): Final Report, Part I". Deliverable of ESPRIT project No. 20268, 1996.

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

[3] P.F.Brown et al. "A Statistical Approach to Machine Translation". *Comp. Linguistics*, 16, 2, pp.79-85, 1990.

[4] A.Castellanos, E.Vidal, J.Oncina. "Language Understanding and Subsequential Transducer Learning". *1st International Colloquium on Grammatical Inference*, Colchester, England. proc., pp. 11/1-11/10. April, 1993.

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

[6] J.A. Feldman, G. Lakoff, A. Stolcke, S.H. Weber. "Miniature Language Acquisition: A touchstone for cognitive science". Tech. Rep., TR-90-009. ICSI, Berkeley, 1990.

[7] R. Haeb-Umbach, H. Ney: "Improvements in Time-synchronous Beam-Search for 10000-word continuous speech recognition". *IEEE Transaction on Speech and Audio Processing.* Vol 2. pp. 353-356. (1994).

[8] F. Jelinek: "Language Modeling for Speech Recognition". In [10] (1996).

[9] V.M. Jiménez, A. Castellanos and E. Vidal: "Some Results with a Trainable Speech Translation and Understanding System", *ICASSP-95, Proc.*, Detroit, 1995.

[10] A.Kornai (ed.); *Proceedings of the ECAI'96 Workshop: Extended Finite State Models of Language.* Budapest, 1996.

[11] J.Oncina, P.Garcia, E.Vidal. "Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks". *IEEE Trans. on PAMI*, Vol.15, No.5, pp.448-458. 1993.

[12] J.Oncina, A.Castellanos, E.Vidal, V.Jimenez. "Corpus-Based Machine Translation through Subsequential Transducers". *Third Int. Conf. on the Cognitive Science of Natural Language Processing*, proc., Dublin, 1994

[13] J.Oncina, M.Varo: "Using Domain Information During the Learning of a Subsequential Transducer". In *Grammatical Inference: Learning Syntax from Sentences*, L.Miclet, C.De La Higuera, Eds. LNAI (1147), Springer Verlag, 1996.

[14] R. Pieraccini, E. Levin, E. Vidal. "Learning How To Understand Language". EUROSPEECH'93, proc., Vol.2, pp. 1407-1412. Berlin, 1993.

[15] K.Seymore, R.Rosenfeld. "Scalable Backoff Language Models". *ICSLP-96, proc.*. pp.232-235. Philadelfia, 1996.

[16] E. Vidal: "Language Learning, Understanding and Translation", In *Proc. in Art. Intell.: CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology*, H. Niemann, R. de Mori and G. Hanrieder (eds.), pp. 131–140. Infix, (1994).

[17] E.Vidal, F.Casacuberta, P.Garcia. "Grammatical Inference and Automatic Speech Recognition". In *Speech Recognition and Coding. New Advances and Trends*, J.Rubio and J.M.Lopez, Eds. Springer Verlag, 1994.

[18] E.Vidal, D.Llorens. "Using knowledge to improve N-Gram Language Modelling through the MGGI methodology". In *Grammatical Inference: Learning Syntax from Sentences*, L.Miclet, C.De La Higuera, Eds. LNAI (1147), Springer-Verlag, 1996.

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

[20] J.M. Vilar, D. Llorens, E. Vidal: "Experiments with Finite-State Models for Speech-Input Language Translation", *Proc. of the SPECOM-96*, St-Petersburg, (1996)

[21] J.M. Vilar, E. Vidal and J.C. Amengual: "Learning Extended Finite State Models for Language Translation". In [10] (1996).